GDACertPrep.com

The Google Data Analytics Certificate Study Guide

Your complete prep guide — 8 chapters, 100 practice questions, SQL reference, portfolio templates, and a full annex.

8 Chapters

100 Q&As

SQL Cheat Sheet

Portfolio Templates

PREMIUM EDITION 2026 · GDACertPrep.com · $29 VALUE

GDACertPrep.com

The Google Data Analytics Certificate Study Guide

Your complete prep guide — 8 chapters, 100 practice questions, SQL reference, portfolio templates, and a full annex.

8 Chapters

100 Q&As

SQL Cheat Sheet

Portfolio Templates

PREMIUM EDITION 2026 · GDACertPrep.com · $29 VALUE

Table of Contents

How to Use This Guide

Passing the Google Data Analytics Professional Certificate is completely achievable — but only if you study smart. This guide was built to be the companion you wish you had from day one: a single resource that explains the concepts clearly, shows you exactly what's tested, and gives you realistic practice before every graded assessment.

About the Google Data Analytics Certificate

The Google Data Analytics Professional Certificate is a self-paced program offered on Coursera. It consists of **8 courses** covering the full data analytics workflow — from asking the right questions, to cleaning and analyzing data, to sharing insights through visualization. The program teaches four primary tools: **Google Sheets, SQL (BigQuery), Tableau,** and **R programming**. It is designed to take approximately **6 months** at 10 hours per week, though many students complete it faster. Upon completion, you earn a certificate backed by Google and recognized by 150+ employer partners including Deloitte, Target, Verizon, and Wells Fargo.

Each of the 8 courses contains multiple modules with videos, readings, hands-on activities, and graded quizzes. You must score at least **80%** on each graded quiz to pass — and you can retake quizzes if needed. The program culminates in a **capstone project** (Course 8) where you complete a real-world case study and build the foundation of your professional portfolio.

How This Study Guide Is Organized

- **Chapters 1–8** mirror the 8 Coursera courses exactly. Each chapter covers the key concepts, vocabulary, and skills from that course.
- **Practice Questions** at the end of each chapter reflect the style and difficulty of real Coursera quiz questions. Work through them without looking at the answers first.
- **Answer Key** (after Chapter 8) provides complete answers with explanations for all 100 questions.
- **SQL Quick Reference** is a condensed cheat sheet covering every SQL concept tested in the program.
- **Spreadsheet Formula Cheat Sheet** covers the essential Google Sheets functions you need to know cold.
- **Annex** includes a full glossary, job market data, recommended resources, and next steps after passing.

Recommended Study Schedule

Week	Focus	This Guide
1	Course 1: Foundations	Read Chapter 01 · Complete practice questions
2	Course 2: Ask Questions	Read Chapter 02 · Review spreadsheet formulas cheat sheet

3	Course 3: Prepare Data	Read Chapter 03 · Study SQL intro section
4	Course 4: Process Data	Read Chapter 04 · Practice SQL cleaning queries
5	Course 5: Analyze Data	Read Chapter 05 · Master all SQL examples
6	Course 6: Share/Visualize	Read Chapter 06 · Set up Tableau Public
7	Course 7: R Programming	Read Chapter 07 · Run R code examples in RStudio
8	Course 8: Capstone	Read Chapter 08 · Use portfolio templates
9	Review Week	Re-do all practice questions · Study answer key explanations
10	Final Push	Review annex · SQL cheat sheet · Take Coursera assessments

■ PRO TIP

Use this guide alongside the Coursera videos — not instead of them. Watch the lesson first, then read the corresponding section here to reinforce and deepen what you learned. The practice questions work best when you treat them like real quiz conditions: no peeking at the answers until you've committed to a choice.

Tips for Studying Effectively

- **Don't rush quizzes.** Read every answer option carefully. The Coursera quizzes often include "almost right" distractors that test whether you truly understand the concept.
- **Code along.** For Chapters 03–05 (SQL) and Chapter 07 (R), actually run the example queries and code. BigQuery Sandbox and RStudio Cloud are both free.
- **Build your glossary early.** Data analytics has a lot of precise vocabulary. Use the Vocab callout boxes in each chapter to build your personal reference list.
- **Use the Annex.** The annex contains a comprehensive glossary and exam-mistake guide that many students find invaluable in the final week before their assessments.
- **Start your portfolio in Week 8.** The capstone case study takes longer than most students expect. Use the portfolio templates in Chapter 08 to structure your write-up.

CHAPTER 01

Foundations of Data Analytics

The data analysis process · data-driven thinking · key roles · stakeholders

14 pages · 10 practice questions

Introduction: What Data Analytics Really Is

Every business decision used to be made on gut instinct. A store manager would order more inventory because it "felt like" demand was rising. A marketing team would run the same ads they'd always run because they worked before. That era is over. Today, data is the foundation of nearly every strategic decision — and the people who can collect, clean, analyze, and communicate data insights are among the most sought-after professionals in every industry.

The Google Data Analytics Certificate, offered through Coursera, is designed to give you the foundational skills to become a junior data analyst. The program teaches the full analytical workflow — from asking smart questions to presenting polished visualizations — using tools like SQL, Google Sheets, Tableau, and R. This first course establishes the conceptual foundation everything else builds on.

1.1 What Is Data Analytics?

Data analytics is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making. It's important to distinguish between the related fields you'll encounter:

Role	Focus	Key Tools
Data Analyst	Analyzes existing data to answer business questions	SQL, Sheets, Tableau, R
Data Scientist	Builds models to make predictions from data	Python, ML frameworks, Statistics
Data Engineer	Builds and maintains pipelines that move data	Cloud platforms, ETL tools
Business Analyst	Translates business needs into data requirements	Excel, BI tools, SQL

■ VOCABULARY: Data Analytics

The science of analyzing raw data to find trends, answer questions, and draw conclusions that support better decision-making.

■ VOCABULARY: Data Ecosystem

The various elements that interact to produce, manage, store, organize, analyze, and share data — including hardware, software, people, and processes.

1.2 The Six-Step Data Analysis Process

The heart of Course 1 — and the framework that runs through the entire certificate — is the six-phase data analysis process. Memorize these phases and what happens in each one.

ASK	Define the problem. Ask SMART questions. Identify stakeholders and understand their expectations. What decision needs to be made? What data do we need?
PREPARE	Collect and store the data. Consider data types, formats, and sources. Assess bias and credibility. Organize files following best practices.
PROCESS	Clean the data. Handle null values, remove duplicates, fix errors. Document every change in a changelog. Verify the cleaned data.

ANALYZE	Explore, transform, and model the data. Use SQL queries, spreadsheet functions, pivot tables, and R to find patterns and answer the original question.
SHARE	Create visualizations and presentations. Use Tableau, R, or Google Slides to present insights clearly to stakeholders.
ACT	Apply the insights to make business decisions. Document recommendations and plan for follow-up analysis.

■ PRO TIP

These six phases appear in quiz questions throughout the entire certificate — not just Course 1. When a scenario describes an analyst doing something, ask yourself: which phase is this? Expect at least 2–3 questions on this in every course.

1.3 The Data Life Cycle

Distinct from the analysis process, the **data life cycle** describes how data is managed from creation through deletion. Understanding this helps you know when data is appropriate to use and how to handle it responsibly.

- **Plan:** Decide what kind of data is needed, how it will be managed, and who will be responsible for it.
- **Capture:** Collect or receive data from various sources — surveys, sensors, transactions, databases.
- **Manage:** Maintain and care for data, determining how and where it's stored and how it's kept secure.
- **Analyze:** Use data to solve problems, make decisions, and support business goals.
- **Archive:** Store data for long-term access in a different location from current storage.
- **Destroy:** Remove data using secure methods, especially when it contains private information.

■ VOCABULARY: Data Life Cycle

The sequence of stages data goes through, from creation and collection to archival and destruction. Different from the data analysis process, which is about what analysts do with data.

■ **WATCH OUT**

The data life cycle and the data analysis process are frequently confused on quizzes. Remember: the life cycle is about data management and governance. The analysis process is about what the analyst does to extract insights. They have different phases with different names.

1.4 Data-Driven vs. Data-Inspired Thinking

One of the key distinctions introduced in Course 1 is the difference between **data-driven** and **data-inspired** decision-making.

Data-driven decision-making means using the facts, metrics, and evidence gathered through data analysis to guide business strategy. The decision follows directly from what the data shows. For example: "Our A/B test showed a 12% higher conversion rate for Version B, so we'll roll out Version B to all users."

Data-inspired decision-making means exploring different data sources to find out what they have in common and use those insights to guide the decision-making process. The data informs but doesn't dictate the decision. Creative, qualitative, and experiential factors also play a role.

■ **VOCABULARY: Data-Driven Decision-Making**

Using facts, metrics, and data to guide strategic business decisions, with the data directly determining or strongly informing the outcome.

Both approaches are legitimate. Analysts should understand which approach their organization favors and what role the data plays in the final decision.

1.5 Key Roles in a Data Team

Data analytics doesn't happen in isolation. A full data ecosystem involves multiple specialized roles that work together. As a data analyst, you'll interact with all of them.

- **Data Analyst:** Collects, cleans, and analyzes data to answer specific questions. Uses SQL, spreadsheets, and visualization tools. Entry-level to mid-level role.
- **Data Scientist:** Uses advanced statistical methods and machine learning to predict future outcomes. Typically requires a stronger programming background (Python, R).
- **Data Engineer:** Builds the infrastructure (pipelines, databases, warehouses) that makes data available for analysts and scientists.
- **Business Intelligence (BI) Analyst/Developer:** Creates dashboards and reports that help businesses track KPIs and operational metrics in real time.

- **Database Administrator (DBA):** Manages databases — including security, performance, backups, and user access.
- **Statistician:** Uses statistical theory and methods to collect, analyze, and interpret data, often in research contexts.

■ PRO TIP

On Coursera quizzes, you'll be given a description of what someone does at work and asked to identify their role. Focus on the *outputs*: analysts produce insights and reports, scientists build predictive models, engineers build data infrastructure.

1.6 Analytical Thinking Skills

Google identifies five core analytical thinking skills that effective data analysts use. These aren't just soft skills — they define how you approach problems:

Skill	Definition	Example
Curiosity	Seeking new challenges and experiences	Wondering why a metric suddenly dropped
Understanding Context	Putting information into a bigger framework	Knowing that Q4 spikes are seasonal
Technical Mindset	Breaking down processes into smaller steps	Structuring a cleaning process step-by-step
Data Design	Organizing information logically	Creating a clear database schema
Data Strategy	Managing processes and tools to achieve goals	Choosing the right tool for each analysis task

1.7 Understanding Stakeholders

Stakeholders are people who have invested interest in a project's outcome. Knowing who your stakeholders are and what they care about is essential to delivering analysis that actually gets used.

- **Executive stakeholders** (C-suite, VPs): Want high-level summaries and strategic insights. Don't need every technical detail — they need the "so what."
- **Project stakeholders** (managers, team leads): Need enough detail to make decisions and allocate resources. Will ask follow-up questions.
- **Analyst team members:** May need full technical depth so they can verify your work or extend it.
- **Customers/end users:** In some projects, the people your analysis affects are also stakeholders whose needs should be considered.

■ **VOCABULARY: Stakeholder**

Anyone who has a vested interest in the outcome of a project — whether they're funding it, executing it, or affected by its results.

A key skill introduced in Course 1 (and tested throughout the certificate) is **stakeholder management** — understanding what each stakeholder needs, setting clear expectations, and communicating your findings in a way that's appropriate for their role and expertise level.

■ **WATCH OUT**

A common quiz mistake: confusing "primary stakeholder" with "most important person." In the GDA certificate, primary stakeholders are those who are directly affected by or have the highest interest in the project outcome. Secondary stakeholders are indirectly affected.

1.8 The Tools of Data Analytics

The certificate introduces four primary tools. You'll use all of them across the 8 courses. Here's a quick overview of what each is used for:

Tool	Best For	Covered In
Google Sheets / Excel	Small-to-medium datasets, quick calculations, pivot tables, VLOOKUP	Courses 2, 4, 5
SQL (BigQuery)	Large datasets stored in relational databases, joining multiple tables	Courses 3, 4, 5
Tableau (Public)	Interactive data visualizations and dashboards	Course 6
R (RStudio)	Statistical analysis, advanced visualization, reproducible reports	Course 7

■ **PRO TIP**

You don't need to master all four tools before starting. The courses introduce each tool progressively. That said, getting comfortable with Google Sheets early (Course 2) makes Courses 3–5 significantly easier.

■ WHAT TO REMEMBER

- The six phases of data analysis are: Ask, Prepare, Process, Analyze, Share, Act.

- The data life cycle (Plan, Capture, Manage, Analyze, Archive, Destroy) is different from the analysis process.

- Data-driven decisions follow directly from data. Data-inspired decisions use data as one input among several.

- Key roles: data analyst (insights), data scientist (predictions), data engineer (infrastructure).

- Five analytical thinking skills: curiosity, understanding context, technical mindset, data design, data strategy.

- Stakeholders can be primary (directly affected) or secondary (indirectly affected).

- The four main tools: Google Sheets, SQL/BigQuery, Tableau, and R.

Chapter 01 — Practice Questions

Answer all questions before checking the Answer Key at the back of this guide.

Q1. Which of the following BEST describes the difference between the data life cycle and the data analysis process?

A) They are the same thing described in different ways

B) The data life cycle focuses on data management; the analysis process focuses on what analysts do with data

C) The data analysis process has six phases; the data life cycle has only three

D) The data life cycle is used by engineers; the analysis process is used by scientists

Q2. A data analyst at a retail company notices that sales drop every January. Rather than running a formal analysis, the executive team decides to increase marketing spend in February based on this observation and past experience. This is an example of:

A) Data-driven decision-making

B) Data-inspired decision-making

C) Qualitative analysis

D) Confirmation bias

Q3. During which phase of the data analysis process does an analyst clean dirty data, handle null values, and remove duplicates?

A) Ask

B) Prepare

C) Process

D) Analyze

Q4. A company's VP of Marketing wants a one-page summary of last quarter's campaign results with three key takeaways. Which stakeholder category does this VP represent?

A) Secondary stakeholder

B) Technical stakeholder

C) Executive stakeholder

D) Project team member

Q5. Which role is MOST focused on building and maintaining the data pipelines and infrastructure that make data available to analysts?

A) Data Analyst

B) Data Scientist

C) Database Administrator

D) Data Engineer

Q6. An analyst creates a logical schema for storing customer purchase records so that the data is easy to query and join. This skill is BEST described as:

A) Curiosity

B) Data strategy

C) Data design

D) Understanding context

Q7. Which of the following CORRECTLY matches a tool to its primary use in the GDA certificate program?

A) Tableau — writing SQL queries to clean large datasets

B) R — building interactive dashboards for executive stakeholders

C) BigQuery — querying large relational databases using SQL

D) Google Sheets — building machine learning prediction models

Q8. A company collects customer data but is legally required to delete it after three years. Which phase of the data life cycle does the deletion represent?

A) Archive

B) Manage

C) Destroy

D) Capture

Q9. Data that is generated by measuring something physical — like temperature readings from a sensor — is collected during which phase of the data life cycle?

A) Plan

B) Capture

C) Analyze

D) Manage

Q10. A senior analyst reviews a newly hired analyst's work and finds their conclusions are supported by the data but ignore a key seasonal trend that explains the pattern. Which analytical thinking skill is the new analyst missing?

A) Technical mindset

B) Data strategy

C) Understanding context

D) Curiosity

CHAPTER 02

Ask Questions to Make Data-Driven Decisions

SMART questions · structured thinking · spreadsheet basics · stakeholder communication

14 pages · 12 practice questions

Introduction: Why Asking the Right Question Is Half the Work

Here's something that surprises many new analysts: the most common reason a data project fails isn't bad data or wrong analysis — it's asking the wrong question in the first place. If you ask "How many customers visited our website this month?" you'll get an answer. But if the real business problem is "Why are customers leaving without buying?" you've answered the wrong question entirely and wasted everyone's time.

Course 2 of the Google Data Analytics Certificate is built around this insight. Before touching a single spreadsheet cell or writing a single SQL query, a great analyst asks the right questions — questions that are specific, measurable, and directly tied to the problem they're trying to solve. This chapter covers the SMART question framework, structured thinking, and the spreadsheet fundamentals you'll use throughout the program.

2.1 The SMART Questions Framework

The SMART framework gives you a structured way to evaluate whether a question is good enough to drive useful analysis. Each letter stands for a quality that effective questions should have:

Letter	Stands For	What It Means	Example
S	Specific	Simple, significant, and focused on a single topic	"What was the average order value for mobile users in Q3?"
M	Measurable	Has specific criteria for success; can be quantified	"Which product category had the highest return rate?"
A	Action-oriented	Encourages change and drives a concrete next step	"What changes to checkout flow would reduce cart abandonm
R	Relevant	Matters to the problem and is worth solving	"How does delivery time affect repeat purchase rate?"
T	Time-bound	Specifies a time frame or date range	"How did revenue change in the 30 days after the campaign la

■ VOCABULARY: SMART Question

A question that is Specific, Measurable, Action-oriented, Relevant, and Time-bound. SMART questions guide analysis toward clear, actionable insights.

Understanding SMART questions also means recognizing what makes a question **not** SMART. Vague or leading questions are two common problems:

- **Vague questions** are too broad to guide analysis: "How is our business doing?" has no clear answer because "doing" isn't defined.
- **Leading questions** suggest a particular answer and introduce bias: "Don't you think customers prefer our new design?" pushes toward a yes.
- **Closed-ended questions** generate only yes/no responses when open-ended questions would yield richer data.
- **Unfair questions** make assumptions about the respondent: "Why did you switch to our competitor?" assumes they did.

■ WATCH OUT

Coursera quiz questions often present four versions of a question and ask which one is SMART. The trick: eliminate options that are vague ("how is performance?"), leading ("don't you agree that..."), or missing a time element ("what are our sales?" vs. "what were our Q3 sales compared to Q2?").

2.2 Types of Data Questions

Beyond the SMART framework, analysts work with several types of questions that serve different analytical purposes. Knowing which type of question a stakeholder is actually asking helps you figure out the right approach:

- **Descriptive questions** ask "What happened?" — they describe past events. Example: "How many units did we sell last month?"
- **Diagnostic questions** ask "Why did it happen?" — they explain causes. Example: "Why did sales drop in the northeast region?"
- **Predictive questions** ask "What will happen?" — they forecast. Example: "Which customers are most likely to churn in the next 30 days?"
- **Prescriptive questions** ask "What should we do?" — they recommend actions. Example: "What discount offer would most effectively retain at-risk customers?"

■ **PRO TIP**

Most of the work in the GDA certificate focuses on **descriptive** and **diagnostic** questions — the foundational layer of analytics. Data science and machine learning courses move into predictive and prescriptive territory.

2.3 Structured Thinking

Structured thinking is the process of recognizing the current problem or situation, organizing available information, revealing gaps and opportunities, and identifying options to solve the problem. It's the mental discipline that separates analysts who deliver clear, actionable insights from those who produce confusing reports.

The GDA certificate introduces a **problem-solving framework** that applies structured thinking to real business scenarios:

1. Define the problem	State clearly what the analysis is trying to solve. Make sure everyone agrees on the problem before collecting any data.
2. Gather context	Understand the background: industry, company goals, past performance, known constraints.
3. Identify available data	What data already exists? What needs to be collected? What are the gaps?

4. Choose the right tools	Will this be solved with a pivot table, a SQL query, or a full visualization dashboard?
5. Communicate findings	Structure your output for your audience — executive summary for leadership, detailed report for your team.

■ VOCABULARY: Structured Thinking

The analytical process of recognizing a problem, organizing available information, identifying gaps, and determining a clear path to a solution.

2.4 Spreadsheet Basics

Spreadsheets are the data analyst's Swiss army knife for small-to-medium datasets. Google Sheets (used in this program) and Microsoft Excel share nearly identical functionality for the purposes of this certificate. You need to be comfortable with three core skills: sorting, filtering, and basic formulas.

Sorting Data

Sorting organizes data rows in ascending or descending order based on one or more columns. In Google Sheets, go to **Data > Sort range**. Key options:

- **Sort by a single column:** Data > Sort sheet by column A, A→Z (ascending) or Z→A (descending).
- **Sort by multiple columns:** Data > Sort range > Add another sort column. Useful for "sort by region, then by revenue."
- **Sort with a header row:** Always check "Data has header row" to prevent your header from being sorted into the data.

Filtering Data

Filtering hides rows that don't match your criteria — the data is still there, just not visible. This is useful for focusing on a subset without deleting anything.

- Enable filters: Data > Create a filter. Filter icons (▼) appear in the header row.
- Click the filter icon to filter by specific values, conditions (greater than, contains, etc.), or a custom formula.
- Remove filters: Data > Remove filter. All hidden rows reappear.

Essential Formulas for Course 2

Formula	What It Does	Example
`=SUM(range)`	Adds all values in a range	=SUM(B2:B50)

`=AVERAGE(range)`	Calculates the mean of a range	=AVERAGE(C2:C100)
`=COUNT(range)`	Counts cells with numbers	=COUNT(A2:A500)
`=COUNTA(range)`	Counts non-empty cells (including text)	=COUNTA(D2:D200)
`=MAX(range)`	Returns the largest value	=MAX(E2:E50)
`=MIN(range)`	Returns the smallest value	=MIN(E2:E50)

■ PRO TIP

Learn the keyboard shortcuts early. In Google Sheets: **Ctrl+Shift+L** (or Cmd+Shift+L on Mac) adds/removes filters. **Ctrl+Home** jumps to cell A1. **Ctrl+End** jumps to the last data cell. These will save you significant time when working with large datasets.

2.5 Communicating with Stakeholders

Even the best analysis is worthless if you can't communicate it clearly. Course 2 introduces key principles for effective stakeholder communication — skills that apply throughout the entire certificate and your career.

Know Your Audience

Different stakeholders need different types of communication. An executive needs a headline and three bullet points. A fellow analyst needs the methodology and raw data access. A product manager needs actionable recommendations tied to timelines. Tailor every deliverable to who will receive it.

Lead with the Insight, Not the Data

A common beginner mistake is to present data first, then conclusions. Busy stakeholders want the answer first: "Sales are down 18% YoY in the northeast. Here's why." Then support that claim with data. This structure — called **BLUF (Bottom Line Up Front)** — is used by everyone from Google analysts to military commanders because it works.

Set Clear Expectations

Before starting any project, align with stakeholders on: What questions are you answering? What will the deliverable look like? When will it be ready? What data sources will you use? This prevents the frustrating situation where you deliver a thorough analysis and the stakeholder says "That's not what I was expecting."

Handling Conflicts and Pushback

Sometimes stakeholders don't like what the data shows. Maybe it challenges a decision they've already made, or contradicts their intuition. Your job is not to make the data say what they want — it's to present the data honestly and explain your methodology clearly. Good analysts separate the data from the recommendation and remain objective.

■ **VOCABULARY: Stakeholder Communication**

The process of sharing data findings, progress updates, and recommendations with the people who have a stake in the project's outcome, tailored to their role, expertise, and information needs.

■ **WATCH OUT**

Never change or omit data to please a stakeholder. This is an ethical violation and can have serious professional and legal consequences. If stakeholders push back on your findings, acknowledge their perspective and walk them through your methodology transparently — but don't alter the analysis to show a different result.

2.6 Common Communication Formats

- **Email updates:** Brief status updates with a subject line that summarizes the key point. Use bullet points. Attach any data files.
- **Meeting presentations:** Slides with clear titles that state the conclusion (not just the topic). Lead with the executive summary slide.
- **Written reports:** Detailed documents for technical or regulatory audiences. Include methodology, data sources, and limitations.
- **Dashboards:** Real-time views of key metrics for stakeholders who need ongoing visibility, not one-time analysis.

■ WHAT TO REMEMBER

- SMART questions are Specific, Measurable, Action-oriented, Relevant, and Time-bound.

- Avoid vague, leading, closed-ended, or unfair questions — they produce bad analysis.

- The four question types: descriptive (what), diagnostic (why), predictive (will), prescriptive (should).

- Structured thinking: define the problem → gather context → identify data → choose tools → communicate.

- Spreadsheet basics: sort (organize rows), filter (hide non-matching rows), formulas (calculate values).

- Always tailor communication to your audience. Lead with the insight (BLUF), not the data.

- Never alter data or analysis to satisfy stakeholder preferences.

Chapter 02 — Practice Questions

Answer all 12 questions before checking the Answer Key.

Q1. A product manager asks: 'Is our app performing well?' Which quality of a SMART question is this MOST lacking?

A) Measurable — there's no way to quantify 'performing well'
B) Action-oriented — it doesn't encourage a change
C) Specific — the question is too vague to guide analysis
D) Both A and C — it is neither specific nor measurable

Q2. An analyst is asked: 'Did customers who received the new email campaign spend more than those who didn't, during the 60-day period after launch?' Which characteristic does this question demonstrate?

A) Leading — it assumes the campaign worked
B) SMART — it is specific, measurable, and time-bound
C) Vague — it doesn't define 'spend more'
D) Unfair — it makes assumptions about customer behavior

Q3. A stakeholder asks: 'Which of our marketing channels will generate the most leads next quarter?' This is BEST classified as which type of analytical question?

A) Descriptive

B) Diagnostic

C) Predictive

D) Prescriptive

Q4. An analyst has collected sales data. The first thing they should do using structured thinking is:

A) Run a pivot table to identify the top-selling products

B) Clearly define the problem they are trying to solve

C) Choose the visualization type for the final report

D) Gather context about the industry

Q5. A data analyst uses a filter in Google Sheets to show only rows where the Region column equals 'West'. What happens to the rows that are filtered out?

A) They are permanently deleted from the spreadsheet

B) They are moved to a new tab automatically

C) They are hidden from view but remain in the data

D) They are highlighted in red to indicate exclusion

Q6. Which formula would you use in Google Sheets to calculate the total revenue in cells B2 through B500?

A) =TOTAL(B2:B500)

B) =ADD(B2,B500)

C) =SUM(B2:B500)

D) =COUNT(B2:B500)

Q7. An analyst is presenting findings to the VP of Sales. According to best practices for stakeholder communication, what should come first in the presentation?

A) The raw data tables so stakeholders can verify the numbers

B) The methodology section explaining how the analysis was conducted

C) The key insight or recommendation (bottom line up front)

D) A comprehensive list of data sources and limitations

Q8. After completing an analysis, an analyst discovers that the results contradict the CEO's strongly held belief about customer behavior. What is the MOST appropriate action?

A) Adjust the analysis to align with the CEO's expectations

B) Present the findings honestly and walk through the methodology transparently

C) Leave out the contradicting data and only present supporting results

D) Delay the presentation until the CEO is in a receptive mood

Q9. Which of the following BEST describes the difference between sorting and filtering in a spreadsheet?

A) Sorting removes data; filtering reorders it

B) Sorting reorders rows; filtering hides rows that don't match a condition

C) Both sorting and filtering permanently delete non-matching rows

D) Filtering reorders rows; sorting hides rows that don't match a condition

Q10. A team is analyzing whether a new employee training program improved productivity. Before collecting any data, the project manager wants to make sure everyone agrees on what 'improved productivity' means. This step reflects which aspect of structured thinking?

A) Gathering context

B) Identifying available data

C) Choosing the right tools

D) Defining the problem

Q11. An analyst uses =COUNTA(A2:A500) in a Google Sheet. What will this formula return?

A) The sum of all numeric values in the range A2:A500

B) The count of cells containing numbers in A2:A500

C) The count of all non-empty cells in A2:A500, including text

D) The average value across all cells in A2:A500

Q12. Which of the following questions is MOST clearly leading?

A) 'What percentage of customers completed the checkout process in Q3?'

B) 'How do return rates compare between Product A and Product B?'

C) 'Don't you think customers prefer our original app design over the new one?'

D) 'What was the average session duration for mobile users in November?'

CHAPTER 03

Prepare Data for Exploration

Data types · data structures · bias · credibility · metadata · databases · SQL intro

16 pages - 12 practice questions

Introduction: The Foundation of Good Analysis

Garbage in, garbage out. It's one of the oldest rules in computing, and it's never been more true than in data analytics. Before you can analyze anything meaningfully, you need to understand where your data comes from, whether it's reliable, how it's structured, and whether it has any inherent biases that could skew your conclusions. Course 3 is where you build that critical foundation.

This chapter covers data types and structures, the many forms of bias that can corrupt datasets, metadata and database fundamentals, data organization best practices, and your first introduction to SQL — the language you'll use to query databases throughout the rest of the certificate.

3.1 Data Types

Every piece of data has a type. Understanding data types helps you choose the right analysis methods and avoid errors when working in spreadsheets and SQL.

Quantitative vs. Qualitative Data

Type	Definition	Examples	Common Analysis
Quantitative	Numerical data that can be counted or measured	Revenue, temperature, age, units sold	Mean, sum, standard deviation
Qualitative	Descriptive data about qualities or characteristics	Customer feedback, product category, country name	Frequency counts, themes, categorization

Discrete vs. Continuous Data

- **Discrete data** can only take specific, countable values — often whole numbers. You can't have 2.7 customers. Examples: number of orders, headcount, number of product defects.

- **Continuous data** can take any value within a range and can be divided into smaller increments. Examples: temperature (72.4°F), weight (185.3 lbs), time (3.47 seconds).

Nominal vs. Ordinal Data

- **Nominal data** represents categories with no natural order. You can't say one category is "greater than" another. Examples: product color (red, blue, green), country, job title.

- **Ordinal data** has a natural order or ranking, but the intervals between values aren't necessarily equal. Examples: customer satisfaction (1-5 stars), survey responses (strongly agree, agree, neutral, disagree, strongly disagree).

VOCABULARY: Discrete Data

Countable data that can only take specific values, typically whole numbers. You cannot have fractional discrete values.

VOCABULARY: Continuous Data

Measurable data that can take any value within a range, including decimals and fractions.

VOCABULARY: Nominal Data

Categorical data with no natural order or ranking between categories.

VOCABULARY: Ordinal Data

Categorical data with a meaningful order or ranking, though the intervals between ranks may not be equal.

WATCH OUT

A common exam trap: "rating from 1 to 5" sounds quantitative because it uses numbers, but it's actually ordinal. The difference between a 2-star and 3-star rating is not necessarily the same as between a 4-star and 5-star rating. Don't assume numbers automatically mean quantitative/continuous data.

3.2 Data Structures

Beyond types, data also comes in different **structures** — the way it's organized and stored. Understanding structure helps you choose the right tools and approach.

Structure	Description	Examples	Best Analyzed With
Structured	Organized in rows and columns with a defined schema	SQL databases, CSV files, spreadsheets	SQL, spreadsheets, Tableau
Semi-structured	Has some organization but not a strict schema	JSON, XML, email headers	Code, special parsers
Unstructured	No predefined format or organization	Images, audio, video, social media posts, PDFs	NLP, computer vision, manual review

The GDA certificate focuses almost entirely on **structured data** — data that fits neatly into tables with rows and columns. This is the most common format you'll encounter in business analytics contexts, and it's what SQL and spreadsheets are designed to work with.

VOCABULARY: Structured Data

Data organized in a defined format with rows and columns — like a spreadsheet or SQL database table. Each row represents a record; each column represents a field.

3.3 Data Bias and Credibility

One of the most important skills you'll develop as an analyst is the ability to identify and account for **bias** in your data. Bias leads to skewed analysis and wrong conclusions — even when you've done everything else correctly. The GDA certificate covers six types of bias you must know:

Bias Type	Definition	Example
Sampling Bias	Sample is not representative of the full population	Surveying only online customers excludes in-store customers
Observer Bias	Different people interpret the same data differently	Two doctors looking at the same brain scan see different results
Interpretation Bias	Tendency to interpret ambiguous data to confirm existing beliefs	Assuming a sales dip is seasonal when it's actually a product issue

Confirmation Bias	Searching for or favoring information that confirms existing views	An analyst only looking for data that supports their hypothesis
Availability Bias	Using readily available data rather than the most relevant data	Basing analysis on last month's data when multi-year trends are ne
Outlier Bias	Letting extreme values unduly influence results	One celebrity purchase skewing average order value data

PRO TIP

On Coursera quizzes, bias questions typically describe a scenario and ask you to identify the type of bias. The key distinguisher: **sampling bias** is about who's in your dataset (non-representative group). **Confirmation bias** is about the analyst's behavior (seeking only supporting evidence). **Observer bias** is about how different people read the same data differently.

3.4 Data Credibility: The ROCCC Framework

Before using any dataset, ask yourself whether it's credible. The GDA certificate introduces the **ROCCC** framework for evaluating data quality:

R — Reliable	The data is accurate, complete, and unbiased. It can be trusted to represent reality.
O — Original	You are working with first-party data (collected directly from the source), not a copy of a copy.
C — Comprehensive	The dataset includes all critical information needed to answer the question without gaps.
C — Current	The data is up-to-date and relevant to the current time period being analyzed.
C — Cited	The source is clearly documented and credible — you know where the data came from.

3.5 Data Ethics and Privacy

Data ethics is about handling data responsibly, fairly, and legally. As a data analyst, you will regularly encounter data that includes personal or sensitive information. You need to understand the principles that govern how this data should be used.

- **Consent:** Data subjects must knowingly agree to how their data will be used. Collection without consent is unethical (and often illegal).
- **Currency:** Data subjects should have access to information about how their data is being used.

- **Privacy:** Preserving a data subject's information and activity any time a data transaction occurs. Personally Identifiable Information (PII) must be protected.

- **Openness:** Free access, usage, and sharing of data should be supported — but balanced with privacy concerns.

- **Anonymization:** Removing identifying information from datasets so individuals cannot be identified.

VOCABULARY: PII (Personally Identifiable Information)

Any data that could be used to identify a specific individual — including name, address, email, phone number, social security number, and biometric data.

VOCABULARY: Data Anonymization

The process of protecting private or sensitive data by eliminating personally identifying information from a dataset.

WATCH OUT

Never share a dataset that contains PII without proper authorization and anonymization. This includes sharing data files over email or in public repositories. Even partial identifiers (age + zip code + gender) can sometimes be used to re-identify individuals — a process called "re-identification."

3.6 Metadata

Metadata is often defined as "data about data." It's the information that describes your dataset — who created it, when, how, and what it contains. Metadata is essential for understanding data you didn't collect yourself and for maintaining data quality over time.

- **Descriptive metadata:** Identifies and describes a dataset (title, author, date created, subject matter).

- **Structural metadata:** Describes how data is organized (field names, data types, table relationships).

- **Administrative metadata:** Covers data management details (where the data is stored, who has access, how long it's retained).

In practice, you'll encounter metadata in **data dictionaries** — documents that define what each field in a dataset means, what values it can contain, and how it was collected. Always read the data dictionary before starting any analysis on an unfamiliar dataset.

VOCABULARY: Metadata

Data that describes other data. Includes information about a dataset's content, format, structure, creation date, and provenance.

VOCABULARY: Data Dictionary

A document that defines and describes each field in a dataset — including field name, data type, allowable values, and description.

3.7 Relational Databases

Most business data lives in **relational databases** — systems that organize data into tables that can be linked together using shared fields. Understanding the basics of relational databases is essential before writing SQL queries.

- **Table:** A collection of related data organized in rows (records) and columns (fields). Like a spreadsheet tab, but more powerful.
- **Primary Key:** A column (or combination of columns) that uniquely identifies each row in a table. Every table should have one. Example: customer_id.
- **Foreign Key:** A column in one table that references the primary key of another table, creating a relationship between the tables.
- **Schema:** The structural blueprint of a database — which tables exist, what columns they have, and how they relate to each other.

Example: An e-commerce database might have three tables: **Customers** (customer_id, name, email), **Orders** (order_id, customer_id, date, total), and **Products** (product_id, name, price). The customer_id in the Orders table is a foreign key that links each order to a customer.

```
-- Conceptual table relationships:

Customers Table          Orders Table
customer_id (PK)  <---  customer_id (FK)
name                     order_id (PK)
email                    date
                         total
```

VOCABULARY: Primary Key (PK)

A column that uniquely identifies each row in a database table. No two rows can have the same primary key value.

VOCABULARY: Foreign Key (FK)

A column in one table that references the primary key of another table, establishing a relationship between the two tables.

3.8 Introduction to SQL

SQL (Structured Query Language) is the standard language for interacting with relational databases. You use SQL to retrieve, filter, sort, and aggregate data stored in database tables. The GDA certificate uses **Google BigQuery** as its SQL environment, but the core SQL syntax is universal.

Your First SQL Query: SELECT and FROM

Every SQL query starts with a **SELECT** statement that specifies which columns to return, and a **FROM** clause that names the table to query.

```
-- Return all columns from a table
SELECT *
FROM orders;

-- Return specific columns only
SELECT customer_id, order_date, total_amount
FROM orders;
```

Filtering with WHERE

The **WHERE** clause filters rows based on a condition. Only rows where the condition is TRUE are returned.

```sql
-- Return only orders over $100
SELECT customer_id, order_date, total_amount
FROM orders
WHERE total_amount > 100;

-- Filter by text value (use single quotes)
SELECT *
FROM customers
WHERE region = 'West';

-- Multiple conditions with AND / OR
SELECT *
FROM orders
WHERE region = 'West' AND total_amount > 50;
```

Sorting with ORDER BY

```sql
-- Sort results ascending (default)
SELECT customer_id, total_amount
FROM orders
ORDER BY total_amount ASC;

-- Sort results descending (largest first)
SELECT customer_id, total_amount
FROM orders
ORDER BY total_amount DESC;
```

Limiting Results with LIMIT

```sql
-- Return only the top 10 rows
SELECT customer_id, total_amount
FROM orders
ORDER BY total_amount DESC
LIMIT 10;
```

PRO TIP

SQL keywords (SELECT, FROM, WHERE, ORDER BY) are not case-sensitive — select and SELECT work the same way. However, the convention (and best practice) is to write SQL keywords in ALL CAPS and table/column names in lowercase. This makes queries much easier to read and debug.

3.9 Data Organization Best Practices

Good data organization prevents confusion, reduces errors, and makes collaboration easier. The GDA certificate introduces specific best practices for naming files and organizing folders.

File Naming Conventions

- Include the **project name** at the start: AirportCampaign_...
- Include the **creation date** in YYYYMMDD format: ..._20260115_...
- Include a **version number**: ..._V01, ..._V02
- Use **underscores** instead of spaces (spaces cause issues in some systems)
- Avoid special characters: no #, %, !, @, or spaces
- Keep names **descriptive but concise**: AirportCampaign_20260115_Sales_V01.csv

Folder Structure

- Create a separate folder for each project
- Within each project, separate **raw data**, **processed data**, and **analysis outputs**
- Keep a **README** file in each project folder explaining what it contains
- **Archive** completed projects — move them to a separate archive folder rather than deleting

PRO TIP

Date formats in file names: always use YYYYMMDD (year-month-day). This ensures files sort correctly in alphabetical order. Files named with MM-DD-YYYY will sort incorrectly across years.

WHAT TO REMEMBER

- Quantitative data is numerical; qualitative data is descriptive.

- Discrete data is countable (whole values); continuous data is measurable (any value).

- Nominal categories have no order; ordinal categories have a natural ranking.

- Structured data (tables) vs. semi-structured (JSON) vs. unstructured (images, text).

- Six bias types: sampling, observer, interpretation, confirmation, availability, outlier.

- ROCCC: data should be Reliable, Original, Comprehensive, Current, and Cited.

- Metadata describes data. Data dictionaries define fields. Always review these before analysis.

- Relational databases: tables linked by primary keys (PK) and foreign keys (FK).

- Basic SQL: SELECT (columns), FROM (table), WHERE (filter), ORDER BY (sort), LIMIT (restrict rows).

- File naming: ProjectName_YYYYMMDD_Description_V01. No spaces or special characters.

Chapter 03 — Practice Questions

Answer all 12 questions before checking the Answer Key.

Q1. A dataset contains customer ratings for a product on a scale of 1 to 5 stars. Which data type BEST describes these ratings?

A) Quantitative and continuous
B) Quantitative and discrete
C) Qualitative and ordinal
D) Qualitative and nominal

Q2. An analyst is studying employee performance and only interviews workers in one department because those employees were most easily available. What type of bias is MOST present?

A) Observer bias
B) Sampling bias
C) Confirmation bias

D) Interpretation bias

Q3. A data analyst evaluates a dataset and determines it was collected three years ago, before a major industry regulation changed. According to the ROCCC framework, this dataset MOST likely fails on which criterion?

A) Reliable
B) Original
C) Comprehensive
D) Current

Q4. In a relational database, a customer_id column in the Orders table that references the customer_id in the Customers table is called a:

A) Primary key
B) Foreign key
C) Schema key
D) Composite key

Q5. Which SQL clause is used to filter rows based on a condition before returning results?

A) ORDER BY
B) SELECT
C) WHERE
D) LIMIT

Q6. An analyst notices that the product team always highlights data that supports a product launch decision while downplaying data that raises concerns. This is BEST described as:

A) Sampling bias
B) Observer bias
C) Confirmation bias
D) Availability bias

Q7. What does metadata provide in the context of a dataset?

A) The cleaned and processed version of the raw data
B) Statistical summaries of each column in the dataset
C) Descriptive information about the data, such as who collected it and when
D) The primary key values used to link tables in a database

Q8. Which of the following is an example of STRUCTURED data?

A) A collection of customer review essays submitted via email

B) Surveillance camera footage from a retail store

C) A spreadsheet tracking monthly sales by region and product category

D) Audio recordings of customer service calls

Q9. A company uses a file naming convention that includes the project name, date in YYYYMMDD format, and a version number. What is the PRIMARY benefit of using YYYYMMDD date format?

A) It shows the month before the year for easier reading

B) Files sort chronologically when sorted alphabetically

C) It uses fewer characters than other date formats

D) It matches the format required by BigQuery

Q10. Which SQL query would return only the top 5 highest-revenue orders from a table named sales_data, with a revenue column?

A) SELECT * FROM sales_data LIMIT 5 WHERE revenue > 0

B) SELECT * FROM sales_data ORDER BY revenue DESC LIMIT 5

C) SELECT TOP 5 revenue FROM sales_data

D) SELECT * FROM sales_data FILTER revenue DESC LIMIT 5

Q11. Two data analysts independently review the same dataset of patient symptoms and classify different patients as high-risk. This is an example of:

A) Sampling bias

B) Confirmation bias

C) Observer bias

D) Availability bias

Q12. A document that defines each field in a dataset — including the field name, data type, and description — is called a:

A) Metadata schema

B) Data dictionary

C) Database index

D) Structural manifest

CHAPTER 04

Process Data from Dirty to Clean

Data cleaning · spreadsheets · SQL · nulls · duplicates · validation · documentation

16 pages - 14 practice questions

Introduction: Clean Data Is the Analyst's Responsibility

Studies suggest that data scientists and analysts spend between 60% and 80% of their time cleaning data. That might sound tedious — and sometimes it is — but it's also where you prevent disasters. An analysis built on dirty data can lead to multi-million dollar decisions made on completely wrong premises. A data analyst who catches and fixes data quality issues before analysis is not just doing busywork: they're protecting the integrity of everything that follows.

Course 4 of the GDA certificate covers the full data cleaning workflow: identifying dirty data, cleaning it in both Google Sheets and SQL, validating the results, and documenting every change. Each of these steps is tested in the Coursera assessments.

4.1 What Makes Data "Dirty"?

Dirty data is data that is incorrect, incomplete, irrelevant, or improperly formatted. It comes in many forms — and recognizing each type is the first step toward fixing it.

Type of Dirty Data	Description	Example
Duplicate records	Same data appears more than once	Customer ID 1042 appears in two rows with slightly different er
Outdated data	Data that is no longer accurate	Employee records showing positions people left two years ago
Incomplete data	Missing values in fields that should have data	Customer records missing phone numbers or zip codes
Incorrect/inaccurate data	Data that is wrong due to human or system error	Age listed as 450; revenue listed as negative
Inconsistent data	Same information formatted differently across records	Date formats: 2024-01-15 vs 01/15/2024 vs Jan 15 2024
Irrelevant data	Data that doesn't apply to the analysis at hand	International customer records in a US-only regional study
Structural errors	Wrong data type in a column	Text values in a numeric price column

VOCABULARY: Data Integrity

The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle.

VOCABULARY: Null Value

A missing or unknown value in a database or spreadsheet. NULL is not the same as zero or an empty string — it represents the absence of any value.

4.2 Data Cleaning in Google Sheets

Google Sheets provides several built-in tools and functions for cleaning data. These are the most commonly tested in the GDA certificate:

TRIM — Removing Extra Spaces

TRIM removes leading spaces, trailing spaces, and extra spaces between words. This is essential when matching values across tables, since "Smith " and "Smith" look the same to the human eye but are treated as different values by a computer.

```
=TRIM(A2)
-- Input:  "  John Smith  "
-- Output: "John Smith"

-- To trim an entire column:
-- In column B, enter =TRIM(A2) then copy down.
```

CLEAN — Removing Non-Printable Characters

CLEAN removes non-printable characters (special characters that don't display visibly but cause errors in processing). These often appear when data is imported from other systems.

```
=CLEAN(A2)

-- Combine with TRIM for thorough cleaning:
=TRIM(CLEAN(A2))
```

Removing Duplicate Rows

- Select the data range you want to deduplicate.
- Go to **Data > Data cleanup > Remove duplicates**.
- Check "Data has header row" if your first row is a header.
- Select which columns to check for duplicates (all columns = exact duplicate match).
- Click "Remove duplicates" — Sheets will report how many rows were removed.

WATCH OUT

Always work on a copy of your data when removing duplicates or making bulk changes. Remove duplicates cannot be undone in some contexts, and you may inadvertently remove records that look like duplicates but aren't. Save the original data in a separate sheet or file first.

COUNTIF — Finding Duplicates Before Removing

```
-- Flag duplicates: formula returns count of how many times value appears
=COUNTIF($A$2:$A$500, A2)

-- Any value greater than 1 is a duplicate.
-- Sort by this helper column to group duplicates together before reviewing.
```

Find and Replace

- Use **Ctrl+H** (or Cmd+H) to open Find and Replace.
- Standardize inconsistent values: replace "CA", "Calif.", "California" all with "California".
- Use "Match entire cell contents" to avoid accidental partial replacements.
- Use "Match case" when case matters.

Data Validation in Google Sheets

Data validation prevents incorrect data from being entered in the first place. Set it up via **Data > Data validation**:

- **List of items:** Create a dropdown with only allowed values (e.g., "Active", "Inactive", "Pending").
- **Number range:** Restrict a cell to numbers between specific values (e.g., 1 to 100).
- **Date range:** Only allow dates within a valid range.
- **Custom formula:** Use any formula to define validation logic.

4.3 Data Cleaning in SQL

While spreadsheet cleaning is great for smaller datasets, SQL is essential for large datasets in databases. The GDA certificate covers several SQL cleaning techniques.

TRIM in SQL

```
-- Remove leading and trailing spaces
SELECT TRIM(customer_name) AS clean_name
FROM customers;

-- LTRIM removes only leading spaces
-- RTRIM removes only trailing spaces
SELECT LTRIM(RTRIM(customer_name)) AS clean_name
FROM customers;
```

Handling NULL Values

```
-- Find rows with NULL values in a column
SELECT *
FROM orders
WHERE customer_id IS NULL;

-- Replace NULL with a default value using COALESCE
-- Returns the first non-NULL value in the list
SELECT
  order_id,
  COALESCE(discount, 0) AS discount_applied
FROM orders;

-- Use IFNULL (BigQuery / MySQL syntax)
SELECT IFNULL(phone_number, 'Unknown') AS phone
FROM customers;
```

Removing Duplicates with DISTINCT

```
-- Return unique values only (removes duplicate rows)
SELECT DISTINCT customer_id
FROM orders;

-- DISTINCT applies to all selected columns combined
SELECT DISTINCT customer_id, product_category
FROM orders;
```

Type Conversion with CAST

CAST converts a value from one data type to another. This is useful when a numeric column is stored as text, or you need to convert date formats.

```
-- Convert a text column to integer
SELECT CAST(price_text AS INT64) AS price_numeric
FROM products;

-- Convert to float (decimal)
SELECT CAST(units AS FLOAT64) AS units_float
FROM inventory;

-- Convert to date
SELECT CAST(order_date_text AS DATE) AS order_date
FROM orders;
```

CASE Statements for Conditional Cleaning

```
-- Standardize inconsistent category values
SELECT
  order_id,
  CASE
    WHEN region = 'CA'         THEN 'California'
    WHEN region = 'Calif.'     THEN 'California'
    WHEN region = 'california' THEN 'California'
    ELSE region
  END AS region_clean
FROM orders;
```

PRO TIP

When cleaning SQL data, always run a SELECT first to preview what the cleaning operation will return before using it in a final query or UPDATE statement. For example: run SELECT TRIM(name) FROM customers LIMIT 20 before applying the transformation across millions of rows.

4.4 Data Validation After Cleaning

After cleaning, always **validate** that your changes worked correctly and didn't introduce new errors. Validation checks include:

- **Row count check:** How many rows did you have before cleaning? How many after? Unexpected changes indicate a problem.
- **NULL count:** How many NULL values remain in critical columns? Are any unexpected?
- **Unique value check:** Are your categorical columns now standardized? Run SELECT DISTINCT region FROM table to verify.

- **Range checks:** Are numeric values within expected ranges? (e.g., no negative ages)
- **Duplicate check:** Run COUNTIF or a GROUP BY query to confirm duplicates are gone.

```
-- Count rows before and after
SELECT COUNT(*) AS row_count FROM orders;

-- Check for remaining NULLs in key column
SELECT COUNT(*) AS null_count
FROM orders
WHERE customer_id IS NULL;

-- Verify unique values in a column
SELECT DISTINCT status, COUNT(*) AS count
FROM orders
GROUP BY status
ORDER BY count DESC;
```

4.5 Documenting Your Cleaning Process

Documentation is not optional — it's a professional requirement. A **changelog** is a record of every change made to a dataset. It tells future analysts (including future you) what changed, when, why, and by whom.

A basic changelog entry should include:

Field	Example
Date	2026-02-15
Change Made	Removed 847 duplicate rows from orders table based on order_id
Reason	Duplicates introduced during data migration from legacy system
Who Made It	J. Smith (Data Analyst)
Before Count	45,923 rows
After Count	45,076 rows

WATCH OUT

Never edit the original raw data file directly. Keep the raw data intact and apply all cleaning in a copy or a derived table. This allows you to go back to the source if a cleaning step was wrong. The changelog should reference what changed — not replace the original data.

WHAT TO REMEMBER

- Dirty data types: duplicates, outdated, incomplete, incorrect, inconsistent, irrelevant, structural errors.

- TRIM() removes extra spaces; CLEAN() removes non-printable characters. Use =TRIM(CLEAN(A2)) for thorough cleaning.

- Always work on a copy of data — never edit raw data files directly.

- Data validation prevents bad data from being entered in the first place.

- SQL: Use TRIM(), COALESCE(), CAST(), and CASE for cleaning. Use IS NULL / IS NOT NULL for null handling.

- DISTINCT in SQL returns unique rows. Use GROUP BY + COUNT to check for duplicates.

- Always validate after cleaning: check row counts, null counts, unique values, and ranges.

- Document every change in a changelog: what changed, when, why, who made it, before/after counts.

Chapter 04 — Practice Questions

Answer all 14 questions before checking the Answer Key.

Q1. An analyst imports customer data and notices that 'Smith ' and 'Smith' are being treated as two different last names. Which Google Sheets function should be applied FIRST to fix this?

A) =CLEAN()
B) =TRIM()
C) =PROPER()
D) =SUBSTITUTE()

Q2. What does a NULL value in a database column represent?

A) The number zero
B) An empty string with no characters
C) The absence of any value — missing or unknown data
D) A deleted record that was not fully removed

Q3. An analyst needs to find all rows in a customers table where the phone_number column has no data. Which SQL clause is correct?

A) WHERE phone_number = NULL

B) WHERE phone_number IS NULL

C) WHERE phone_number = ''

D) WHERE phone_number = 0

Q4. Which SQL function returns the first non-NULL value from a list of expressions?

A) ISNULL()

B) NULLIF()

C) COALESCE()

D) IFELSE()

Q5. An analyst uses =COUNTIF($A$2:$A$500, A2) in column B of their spreadsheet. A value of 3 in column B means:

A) The value in A2 appears in 3 different ranges

B) The value in column A appears 3 times in the range A2:A500

C) There are 3 empty cells in the range

D) The formula found 3 errors in the column

Q6. Before removing duplicate rows from a dataset, what should an analyst do FIRST?

A) Filter the data to show only duplicate rows

B) Create a backup copy of the original dataset

C) Sort the data alphabetically by the ID column

D) Run a VLOOKUP to confirm which rows are duplicates

Q7. A SQL query uses SELECT DISTINCT customer_id FROM orders. What does this return?

A) All rows including duplicates, sorted by customer_id

B) Only the most recent order for each customer

C) A list of unique customer_id values with no repetition

D) The count of how many times each customer_id appears

Q8. Which SQL function is used to convert a column stored as text into a numeric integer data type?

A) CONVERT()

B) FORMAT()

C) CAST()

D) PARSE()

Q9. An analyst uses a CASE statement in SQL to convert 'CA', 'Calif.', and 'california' all to 'California'. This is an example of:

A) Removing duplicate records

B) Standardizing inconsistent data

C) Handling NULL values

D) Filtering irrelevant data

Q10. Which of the following is a data validation technique in Google Sheets?

A) Using TRIM() to remove spaces before data entry

B) Creating a dropdown list so users can only enter approved values

C) Sorting the data alphabetically after entry

D) Running COUNTIF to identify duplicates

Q11. After cleaning a dataset, an analyst runs SELECT COUNT(*) AS row_count FROM orders and compares it to the pre-cleaning count. This is an example of:

A) Data bias detection

B) Data documentation

C) Post-cleaning data validation

D) Structured thinking

Q12. A changelog entry should include which of the following?

A) The raw data in its original, uncleaned state

B) The date, description of change, reason, who made it, and before/after counts

C) Only the SQL queries used to make the changes

D) A summary of the analysis results derived from the cleaned data

Q13. An analyst needs to clean a column that contains non-printable characters imported from a legacy system. Which Google Sheets function should they use?

A) =TRIM()

B) =PROPER()

C) =CLEAN()

D) =SUBSTITUTE()

Q14. Which of the following BEST describes 'data integrity'?

A) Data that has been fully cleaned and has no NULL values

B) The accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle

C) A dataset that has been validated by at least two different analysts

D) Data stored in a relational database with primary and foreign keys

CHAPTER 05

Analyze Data to Answer Questions

SQL aggregations · JOINs · GROUP BY · subqueries · pivot tables · VLOOKUP

18 pages - 16 practice questions

Introduction: This Is Where the Analysis Happens

After you've asked the right questions, gathered your data, and cleaned it thoroughly, you're finally ready to do what most people think of when they picture "data analysis." Chapter 5 — covering Course 5 of the GDA certificate — is the most SQL-intensive part of the program and consistently ranks as the one students find most challenging. But it's also where the real power of data analytics becomes clear: you can interrogate millions of records, combine data from multiple sources, and surface exactly the answers your stakeholders need.

This chapter covers SQL aggregations, GROUP BY, HAVING, all four JOIN types, subqueries, pivot tables in Google Sheets, VLOOKUP, and conditional functions. Work through every code example and run them yourself in BigQuery Sandbox.

5.1 SQL Aggregation Functions

Aggregation functions perform calculations across multiple rows and return a single summary value. They are the backbone of analytical SQL queries.

Function	What It Does	Example	Returns
COUNT(*)	Counts all rows	COUNT(*)	Total number of rows
COUNT(col)	Counts non-NULL values in a column	COUNT(order_id)	Rows where order_id is not NULL
SUM(col)	Adds all numeric values	SUM(revenue)	Total revenue
AVG(col)	Calculates the arithmetic mean	AVG(order_value)	Average order value
MIN(col)	Returns the smallest value	MIN(price)	Lowest price
MAX(col)	Returns the largest value	MAX(price)	Highest price

```
-- How many orders were placed?
SELECT COUNT(*) AS total_orders
FROM orders;

-- What is the total revenue?
SELECT SUM(order_total) AS total_revenue
FROM orders;

-- What is the average order value?
SELECT AVG(order_total) AS avg_order_value
FROM orders;

-- What are the min and max order values?
SELECT
  MIN(order_total) AS min_order,
  MAX(order_total) AS max_order
FROM orders;
```

VOCABULARY: Aggregate Function

A SQL function that performs a calculation on a set of values and returns a single result. Examples: COUNT, SUM, AVG, MIN, MAX.

5.2 GROUP BY — Aggregating by Category

GROUP BY divides rows into groups based on one or more columns, then applies aggregate functions to each group. This is how you answer questions like "What is the total revenue by region?" or "How many orders did each customer place?"

```
-- Total revenue by region
SELECT
  region,
  SUM(order_total) AS total_revenue
FROM orders
GROUP BY region;

-- Number of orders per customer
SELECT
  customer_id,
  COUNT(*) AS order_count
FROM orders
GROUP BY customer_id
ORDER BY order_count DESC;

-- Average order value by product category
SELECT
  category,
  AVG(order_total) AS avg_order_value,
  COUNT(*) AS num_orders
FROM orders
GROUP BY category
ORDER BY avg_order_value DESC;
```

WATCH OUT

When using GROUP BY, every column in your SELECT clause must either be in the GROUP BY clause OR inside an aggregate function. If you SELECT customer_name, region, SUM(revenue) and GROUP BY customer_name only — this will error. You must GROUP BY customer_name, region OR use an aggregate on region.

5.3 HAVING — Filtering Grouped Results

WHERE filters individual rows *before* grouping. **HAVING** filters grouped results *after* aggregation. This distinction is critical and heavily tested in the GDA certificate.

```
-- WHERE filters BEFORE grouping (individual rows)
SELECT region, SUM(revenue) AS total_revenue
FROM orders
WHERE order_date >= '2025-01-01'   -- Filter rows first
GROUP BY region;

-- HAVING filters AFTER grouping (aggregate results)
SELECT region, SUM(revenue) AS total_revenue
FROM orders
GROUP BY region
HAVING SUM(revenue) > 100000;     -- Filter groups

-- Combine both WHERE and HAVING
SELECT
  region,
  COUNT(*) AS order_count,
  SUM(revenue) AS total_revenue
FROM orders
WHERE order_date >= '2025-01-01'    -- Filter rows first
GROUP BY region
HAVING COUNT(*) > 50                -- Then filter groups
ORDER BY total_revenue DESC;
```

PRO TIP

Memory trick for WHERE vs. HAVING: WHERE comes before GROUP BY in the query (and filters before grouping). HAVING comes after GROUP BY (and filters after grouping). The keyword order in your SQL tells you the order of operations.

5.4 SQL JOINs — Combining Multiple Tables

JOINs are the most powerful — and most confusing — concept in SQL for beginners. They let you combine rows from two or more tables based on a related column. Master these and you'll be ready for almost any real-world data analysis task.

INNER JOIN

Returns **only rows that have matching values in both tables**. If a row in the left table has no match in the right table, it's excluded. If a row in the right table has no match in the left table, it's also excluded.

```
SELECT
  customers.customer_name,
  orders.order_id,
  orders.order_date,
  orders.total_amount
FROM customers
INNER JOIN orders
  ON customers.customer_id = orders.customer_id;

-- Only customers who HAVE placed orders appear.
-- Customers with no orders are excluded.
```

LEFT JOIN (LEFT OUTER JOIN)

Returns **all rows from the left table**, plus matching rows from the right table. If there is no match in the right table, the right-side columns contain NULL.

```
SELECT
  customers.customer_name,
  orders.order_id,
  orders.total_amount
FROM customers
LEFT JOIN orders
  ON customers.customer_id = orders.customer_id;

-- ALL customers appear, even those with no orders.
-- Customers with no orders will have NULL in order_id and total_amount.

-- Find customers who have NEVER placed an order:
SELECT customers.customer_name
FROM customers
LEFT JOIN orders ON customers.customer_id = orders.customer_id
WHERE orders.order_id IS NULL;
```

RIGHT JOIN (RIGHT OUTER JOIN)

Returns **all rows from the right table**, plus matching rows from the left table. Less commonly used — you can always rewrite a RIGHT JOIN as a LEFT JOIN by swapping the table order.

```
SELECT
  customers.customer_name,
  orders.order_id
FROM customers
RIGHT JOIN orders
  ON customers.customer_id = orders.customer_id;

-- ALL orders appear, even if the customer record was deleted.
```

FULL OUTER JOIN

Returns **all rows from both tables**. Where there's no match, NULLs fill in the missing side. Useful for finding records that exist in one table but not the other.

```
SELECT
  customers.customer_name,
  orders.order_id
FROM customers
FULL OUTER JOIN orders
  ON customers.customer_id = orders.customer_id;
```

JOIN Type	Returns	NULLs From
INNER JOIN	Only matching rows from both tables	Neither — non-matches are excluded
LEFT JOIN	All left rows + matching right rows	Right table (unmatched left rows get NULLs on right)
RIGHT JOIN	All right rows + matching left rows	Left table (unmatched right rows get NULLs on left)
FULL OUTER JOIN	All rows from both tables	Both sides for unmatched rows

WATCH OUT

The most commonly missed quiz concept: LEFT JOIN returns NULLs from the RIGHT table for unmatched left rows. Students often get this backwards. Visualize it: LEFT JOIN = keep everything on the left, fill right side with NULLs where there's no match.

5.5 Subqueries

A **subquery** is a query nested inside another query. The inner query (subquery) runs first, and its result is used by the outer query. Subqueries can appear in the SELECT, FROM, or WHERE clause.

```
-- Subquery in WHERE: Find customers who spent above average
SELECT customer_id, total_spent
FROM customers
WHERE total_spent > (
  SELECT AVG(total_spent) FROM customers
);

-- Subquery in FROM: Treat a query result as a table
SELECT
  region,
  AVG(order_total) AS avg_order
FROM (
  SELECT region, order_total
  FROM orders
  WHERE order_date >= '2025-01-01'
) AS recent_orders
GROUP BY region;

-- Subquery in SELECT: Calculated column
SELECT
  customer_id,
  total_spent,
  (SELECT AVG(total_spent) FROM customers) AS overall_avg,
  total_spent - (SELECT AVG(total_spent) FROM customers) AS vs_avg
FROM customers;
```

PRO TIP

When a subquery gets complex, consider using a **Common Table Expression (CTE)** with the WITH keyword instead. CTEs make queries more readable: WITH recent_orders AS (SELECT ... FROM orders WHERE ...) SELECT ... FROM recent_orders. CTEs are not tested heavily in the GDA certificate but are good practice.

5.6 Pivot Tables in Google Sheets

Pivot tables are one of the most powerful analysis tools in Google Sheets. They let you summarize, group, and cross-tabulate data interactively — without writing any formulas. This is also one of the most commonly reported challenging topics among GDA certificate students.

Creating a Pivot Table

- Select your data range (include headers).
- Go to **Insert > Pivot table**. Choose New sheet or Existing sheet.

- The Pivot table editor opens on the right side.
- Drag fields to **Rows**, **Columns**, **Values**, and **Filters**.

Pivot Table Fields Explained

Field Area	What It Does	Example
Rows	Groups data along the vertical axis (left side)	One row per Region
Columns	Groups data along the horizontal axis (top)	One column per Quarter
Values	The data being aggregated (sum, count, avg, etc.)	SUM of Revenue
Filters	Filters the entire pivot table	Show only 2025 data

Example: You have a sales dataset with columns: Region, Quarter, Revenue. To create a pivot table showing total revenue by region and quarter:

- Rows: Region (shows CA, TX, NY, etc. as row labels)
- Columns: Quarter (shows Q1, Q2, Q3, Q4 as column headers)
- Values: SUM of Revenue
- Result: A grid showing revenue for every Region/Quarter combination.

WATCH OUT

The pivot table "Values" field aggregates by SUM by default. If you're counting records, change the aggregation to COUNTA or COUNT. Students often report getting confused when the pivot table is counting values instead of summing them — always check the aggregation type in the Values field editor.

5.7 VLOOKUP

VLOOKUP (Vertical Lookup) searches for a value in the leftmost column of a range and returns a value from a specified column in the same row. It's the go-to function for merging data from two spreadsheet tables.

```
Syntax: =VLOOKUP(search_key, range, index, [is_sorted])

search_key:  The value you're looking up (e.g., a customer ID in cell A2)
range:       The table to search in — must start with the column being searched
index:       Which column number to return from the range (1 = first column)
is_sorted:   FALSE for exact match (almost always use FALSE)

Example:
  In Sheet1, A2 contains customer_id = 1042
  In Sheet2, A:C contains: customer_id | name | email

  =VLOOKUP(A2, Sheet2!$A:$C, 2, FALSE)
  Returns: the customer NAME (column 2 of Sheet2!A:C) for customer_id 1042

  =VLOOKUP(A2, Sheet2!$A:$C, 3, FALSE)
  Returns: the EMAIL (column 3) for the same customer
```

WATCH OUT

The most common VLOOKUP mistake: getting the index number wrong. The index counts from the START of your range — not from column A. If your range is D:F, column D is index 1, E is 2, F is 3. Also: always use FALSE as the last argument unless you're doing an approximate match on sorted data (rare).

5.8 Conditional Functions: SUMIF and COUNTIF

```
-- COUNTIF: Count cells matching a condition
=COUNTIF(range, criteria)

=COUNTIF(B2:B500, "West")          -- Count rows where region = "West"
=COUNTIF(C2:C500, ">100")          -- Count orders where total > 100
=COUNTIF(D2:D500, "Completed")     -- Count completed orders

-- SUMIF: Sum cells where a condition is met
=SUMIF(range, criteria, sum_range)

=SUMIF(B2:B500, "West", C2:C500)   -- Sum revenue for West region only
=SUMIF(D2:D500, "Completed", C2:C500)  -- Sum revenue for completed orders

-- AVERAGEIF: Average where condition is met
=AVERAGEIF(B2:B500, "West", C2:C500)   -- Avg revenue in West region

-- COUNTIFS and SUMIFS accept multiple conditions:
=COUNTIFS(B2:B500, "West", D2:D500, "Completed")
-- Count rows where region=West AND status=Completed
```

WHAT TO REMEMBER

- Aggregate functions: COUNT, SUM, AVG, MIN, MAX — return one value from many rows.

- GROUP BY divides rows into groups for aggregation. Every non-aggregated SELECT column must be in GROUP BY.

- WHERE filters individual rows (before grouping). HAVING filters grouped results (after grouping).

- INNER JOIN: only matching rows from both tables. LEFT JOIN: all left rows + matches from right (NULLs for no match).

- RIGHT JOIN: all right rows + matches from left. FULL OUTER JOIN: all rows from both tables.

- Subqueries nest one query inside another. The inner query runs first.

- Pivot tables: Rows = categories, Columns = secondary categories, Values = aggregated metric.

- VLOOKUP: =VLOOKUP(search_key, range, column_index, FALSE). Index counts from start of range, not column A.

- SUMIF/COUNTIF: apply aggregation only where a condition is met. SUMIFS/COUNTIFS support multiple conditions.

Chapter 05 — Practice Questions

Answer all 16 questions before checking the Answer Key.

Q1. An analyst writes: SELECT region, COUNT(*) FROM orders GROUP BY region. What does this query return?

A) The total revenue for each region

B) A single count of all rows in the orders table

C) The number of orders in each region

D) The regions where order count is greater than zero

Q2. An analyst wants to show only regions with more than 500 orders. Which clause should be used to filter this grouped result?

A) WHERE COUNT(*) > 500

B) FILTER COUNT(*) > 500

C) HAVING COUNT(*) > 500

D) AND COUNT(*) > 500

Q3. A LEFT JOIN between a Customers table and an Orders table is performed. A customer has no orders in the Orders table. What will appear in that customer's row?

A) The row will not appear in the results

B) The order columns will contain NULL values

C) The row will be moved to a separate error table

D) The order columns will contain 0

Q4. An analyst wants to find all customers who have NEVER placed an order. They write a LEFT JOIN between Customers and Orders, then filter results. What condition in the WHERE clause will identify customers with no orders?

A) WHERE orders.order_id = 0

B) WHERE orders.order_id IS NULL

C) WHERE customers.customer_id NOT IN orders

D) WHERE orders.total = NULL

Q5. Which JOIN type returns ALL rows from BOTH tables, with NULLs where there is no match on either side?

A) INNER JOIN

B) LEFT JOIN

C) RIGHT JOIN

D) FULL OUTER JOIN

Q6. In a pivot table, an analyst sets Rows = Product Category, Columns = Quarter, and Values = SUM of Revenue. What does each cell in the pivot table represent?

A) The count of products in that category and quarter

B) The average revenue for that category and quarter

C) The total revenue for that specific product category in that specific quarter

D) The percentage of total revenue for that category

Q7. A sales spreadsheet has customer IDs in column A and region in column B. A second sheet has customer IDs in column A, name in column B, and email in column C. To return the customer NAME using the ID in Sheet1!A2, the correct formula is:

A) =VLOOKUP(A2, Sheet2!A:C, 3, FALSE)

B) =VLOOKUP(A2, Sheet2!A:C, 2, FALSE)

C) =VLOOKUP(A2, Sheet2!B:C, 1, FALSE)

D) =VLOOKUP(A2, Sheet2!A:C, 2, TRUE)

Q8. An analyst writes =SUMIF(B2:B500, 'West', C2:C500). What does this calculate?

A) The count of rows where column B equals 'West'

B) The sum of column C values where the corresponding column B value equals 'West'

C) Whether any value in column C equals 'West'

D) The average of column C values for the West region

Q9. A subquery is placed in the WHERE clause of an outer query. In what order does SQL execute these queries?

A) The outer query runs first, then the subquery uses its results

B) The subquery runs first, then the outer query uses the subquery's result

C) Both queries run simultaneously in parallel

D) The order depends on which table is larger

Q10. Which SQL query correctly finds the average order value for all customers who spent more than the overall average?

A) SELECT AVG(total) FROM orders WHERE total > AVG(total)

B) SELECT AVG(total) FROM orders HAVING total > AVG(total)

C) SELECT AVG(total) FROM orders WHERE total > (SELECT AVG(total) FROM orders)

D) SELECT AVG(total) FROM orders WHERE total > ALL(SELECT total FROM orders)

Q11. An analyst writes SELECT DISTINCT product_category FROM orders GROUP BY product_category. What is wrong with this query?

A) Nothing — DISTINCT and GROUP BY can be combined this way

B) DISTINCT is unnecessary when using GROUP BY, which already returns unique values

C) GROUP BY cannot be used without an aggregate function in the SELECT clause

D) DISTINCT must come before FROM, not before the column name

Q12. In a Google Sheets pivot table, the 'Filters' area is used to:

A) Change the aggregation type applied to the Values field

B) Limit the data included in the entire pivot table based on field values

C) Sort the pivot table results alphabetically

D) Add calculated fields to the pivot table

Q13. An analyst runs: SELECT category, SUM(revenue) FROM orders WHERE year = 2025 GROUP BY category HAVING SUM(revenue) > 50000 ORDER BY SUM(revenue) DESC. What is the correct order of operations for this query?

A) GROUP BY > WHERE > HAVING > ORDER BY

B) WHERE > GROUP BY > HAVING > ORDER BY

C) HAVING > WHERE > GROUP BY > ORDER BY

D) WHERE > HAVING > GROUP BY > ORDER BY

Q14. A Customers table has 500 rows. An Orders table has 2,000 rows. An INNER JOIN returns 1,800 rows. What can you conclude?

A) 200 customers have no orders

B) 200 orders are linked to customers who don't exist in the Customers table

C) The join produced 1,800 matching customer-order pairs

D) Both B and C are correct conclusions

Q15. Which formula counts the number of cells in D2:D500 that contain the value 'Completed' AND where the corresponding cell in E2:E500 is greater than 100?

A) =COUNTIF(D2:D500, 'Completed', E2:E500, '>100')

B) =COUNTIFS(D2:D500, 'Completed', E2:E500, '>100')

C) =COUNTIF(D2:D500 AND E2:E500, 'Completed', '>100')

D) =COUNT(D2:D500, 'Completed') + COUNT(E2:E500, '>100')

Q16. An analyst wants to create a column showing each customer's total spend minus the overall average spend. They use a subquery in the SELECT clause. What does the subquery in this context return?

A) A different value for each row, based on that row's data

B) A single scalar value (the overall average) that is applied to each row

C) A new table that replaces the original FROM table

D) A filtered subset of rows based on a condition

CHAPTER 06

Share Data Through the Art of Visualization

Visualization principles · chart types · Tableau · dashboards · data storytelling

14 pages - 12 practice questions

Introduction: Data Without a Story Is Just Numbers

You've cleaned the data. You've run the queries. You know what the numbers say. Now comes the part that actually changes decisions: communicating your findings in a way that makes the insight immediate, clear, and compelling. Data visualization is not decoration — it's the final step that determines whether your analysis gets acted on or ignored.

Course 6 of the GDA certificate teaches how to choose the right chart type, apply design principles that make data clear and accessible, build interactive dashboards in Tableau Public, and structure a data-driven narrative that moves stakeholders to action.

6.1 Why Visualization Matters

The human brain processes visual information roughly 60,000 times faster than text. A well-designed chart can communicate in seconds what a table of numbers would take minutes to understand. But poor visualization — the wrong chart type, cluttered design, misleading scales — can actively obscure the truth or lead viewers to wrong conclusions.

- **Pre-attentive attributes** are visual properties that the brain processes automatically before conscious attention: color, size, shape, position, and length. Effective visualizations use these properties intentionally.
- **Data-ink ratio** (Edward Tufte): maximize the proportion of ink used to represent data vs. decorative elements. Remove chartjunk — unnecessary grid lines, 3D effects, and decorative borders that add visual noise without adding information.
- **The 5-second rule:** if a stakeholder can't grasp the main point of a chart within 5 seconds, the chart needs to be redesigned.

VOCABULARY: Data Visualization

The graphical representation of data using charts, graphs, maps, and other visual formats to communicate patterns, trends, and insights clearly and efficiently.

VOCABULARY: Pre-attentive Attributes

Visual properties processed by the brain automatically and before conscious attention — including color, size, position, shape, and length. Used strategically to direct the viewer's eye to the most important data.

6.2 Choosing the Right Chart Type

Choosing the wrong chart type is one of the most common visualization mistakes. The right choice depends on what relationship you're trying to show:

Chart Type	Best For	When to Use
Bar / Column Chart	Comparing values across categories	Sales by region, performance by team
Line Chart	Showing trends over time	Monthly revenue, daily active users
Pie / Donut Chart	Showing composition (parts of a whole)	Market share, budget breakdown (max 5 segments)
Scatter Plot	Showing relationship between two variables	Ad spend vs. revenue, age vs. purchase frequency
Histogram	Showing distribution of a single variable	Order value distribution, age distribution
Heat Map	Showing intensity across two dimensions	Sales by day/hour, geographic density

Bubble Chart	Comparing three variables simultaneously	Revenue (x), growth (y), market size (bubble)
Box Plot	Showing statistical distribution and outliers	Salary ranges by department, test score spread

WATCH OUT

Avoid pie charts with more than 5 segments — they become unreadable. Never use 3D charts: the perspective distorts values and makes accurate comparison impossible. Always start your y-axis at zero for bar charts; truncating the axis exaggerates small differences and misleads viewers.

6.3 Design Principles for Effective Visualizations

Color

- Use color to highlight the most important data, not to decorate.

- Limit your palette to 3-5 colors maximum. More colors create confusion.

- Use **diverging color scales** (e.g., red to blue) for data with a meaningful midpoint (above/below average).

- Use **sequential scales** (light to dark) for ordered data (low to high).

- **Color accessibility:** approximately 8% of men have red-green color blindness. Always test your charts with a color-blind simulator or use patterns in addition to color.

Text and Labels

- Every chart must have a clear, descriptive title that states the insight, not just the topic. "Q3 Revenue 18% Below Forecast" is better than "Q3 Revenue."

- Label axes clearly with units: "$000s", "%, "Days".

- Use direct data labels on charts where possible to eliminate the need for a separate legend.

- Minimize legend use — legends require the reader's eye to travel back and forth.

Layout and Simplicity

- Remove all elements that don't add information: unnecessary grid lines, background images, 3D effects.

- Align elements consistently — misaligned charts look unprofessional and distract from the data.

- Use white space generously. A chart that tries to show everything shows nothing.

PRO TIP

The "squint test": squint at your chart until it's blurry. What do you see? The most visually dominant elements should be the most important data points. If decorative elements dominate, strip them back.

6.4 Tableau Fundamentals

Tableau Public is the free version of Tableau used in the GDA certificate. It connects to data sources (CSV, Excel, Google Sheets, databases) and lets you build interactive visualizations and dashboards with a drag-and-drop interface. No coding required — though understanding the underlying data structure is essential.

Tableau Interface Overview

Area	Purpose
Data pane (left)	Lists all fields from your data source. Dimensions (blue) = categorical. Measures (green) = numeric.
Columns/Rows shelf	Drag fields here to set the x and y axes of your chart.
Marks card	Controls color, size, shape, label, and detail of marks on the chart.
Filters shelf	Drag fields here to filter the view. Filters can be shown as interactive controls.
Show Me panel (right)	Recommends chart types based on the fields you've selected.
Worksheet / Dashboard	Individual chart views vs. assembled dashboard with multiple views.

Key Tableau Concepts

- **Dimensions vs. Measures:** Dimensions are categorical fields (text, dates, booleans) shown in blue. Measures are numeric fields shown in green. Dragging a dimension to Rows creates one row per category value.
- **Aggregation:** When you drag a Measure to the view, Tableau automatically aggregates it (usually SUM). Right-click to change to AVG, COUNT, MIN, MAX, etc.
- **Pills:** The blue/green buttons representing fields in your shelves. Discrete fields (blue pills) create headers. Continuous fields (green pills) create axes.
- **Calculated fields:** Create new fields using formulas. Right-click in the Data pane and select "Create Calculated Field."
- **Filters:** Right-click any field and select "Show Filter" to add an interactive filter control to your view.

VOCABULARY: Dimension

In Tableau, a categorical field (shown in blue) used to group, segment, or label data. Examples: Region, Product Category, Customer Name.

VOCABULARY: Measure

In Tableau, a numeric field (shown in green) that can be aggregated. Examples: Revenue, Order Count, Profit Margin.

6.5 Dashboard Design

A **dashboard** is a collection of visualizations displayed together on a single screen, providing an at-a-glance view of key metrics. The goal of a dashboard is to answer the most important questions at a glance — without requiring the viewer to run queries or dig through data themselves.

Dashboard Design Principles

- **Start with the question, not the data.** What does the viewer need to know? Design the dashboard to answer that question first.
- **Hierarchy:** Put the most important metrics at the top-left (where eyes naturally start). Support details lower and to the right.
- **Consistency:** Use the same colors, fonts, and chart styles throughout. A consistent visual language reduces cognitive load.
- **Interactivity:** Use filters and actions to let viewers explore. But don't add so many filters that the dashboard becomes overwhelming.
- **Context:** Numbers without context are meaningless. Always show a comparison: vs. last period, vs. target, vs. benchmark.

VOCABULARY: Dashboard

A visual display of multiple charts and metrics on a single screen, designed to give stakeholders an at-a-glance view of key performance indicators.

6.6 Data Storytelling: The McCandless Method

Data journalist David McCandless identified four elements that make data visualizations compelling and useful. The GDA certificate calls this the **McCandless Method**:

Element	Definition	In Practice
Information	The raw data itself	Your dataset, numbers, facts
Story	The narrative connecting the data	"Sales declined because of X"
Goal	The specific outcome the visualization serves	Helps executives allocate Q4 budget

Visual Form	The chosen representation (chart type, design)	Line chart showing trend with annotation

The most effective visualizations have all four elements working together. A chart with beautiful visual form but no clear story is art, not analysis. A chart with a clear story but poor visual form will be ignored.

Structuring a Data Presentation

- **Setup:** Establish the context — what's the situation, who are the stakeholders, what question are we answering?
- **Conflict:** Introduce the tension — the problem, gap, or unexpected finding that makes this analysis necessary.
- **Resolution:** Present the insight and recommendation — what does the data show, and what should we do about it?

PRO TIP

In Coursera Course 6 assessments, you'll often be asked to critique a visualization or identify what's wrong with it. Common issues to look for: missing axis labels, truncated y-axis, too many colors, pie chart with too many segments, 3D effects, and chart titles that describe the data instead of the insight.

WHAT TO REMEMBER

- Pre-attentive attributes (color, size, position, shape) direct the viewer's eye to what matters most.

- Chart selection: bar = compare categories, line = trends over time, scatter = relationships, pie = composition (max 5 slices).

- Design principles: maximize data-ink ratio, use color purposefully, start y-axis at zero, use clear titles.

- Tableau: Dimensions (blue/categorical) go in Rows/Columns. Measures (green/numeric) are aggregated in the view.

- Dashboard design: most important metrics top-left, consistent styling, always show context and comparison.

- McCandless Method: Information + Story + Goal + Visual Form = effective visualization.

- Data storytelling structure: Setup (context) → Conflict (problem) → Resolution (insight + recommendation).

Chapter 06 — Practice Questions

Answer all 12 questions before checking the Answer Key.

Q1. An analyst creates a chart showing monthly website visitors over the past two years with a line that shows the overall trend. Which chart type is MOST appropriate for this?

A) Bar chart

B) Pie chart

C) Line chart

D) Scatter plot

Q2. A visualization shows market share for 9 different companies using a pie chart. What is the MAIN problem with this design choice?

A) Pie charts can only show two categories

B) Pie charts with more than 5 segments become difficult to read and compare

C) Market share data should always be shown as a bar chart

D) Pie charts cannot display percentages

Q3. Which of the following chart types is BEST for showing the relationship between advertising spend and revenue across 200 campaigns?

A) Histogram

B) Line chart

C) Scatter plot

D) Heat map

Q4. In Tableau, what is the difference between a Dimension and a Measure?

A) Dimensions are numeric fields; Measures are categorical fields

B) Dimensions are categorical fields (blue); Measures are numeric fields (green) that can be aggregated

C) Dimensions go in the Filters shelf; Measures go in the Rows shelf

D) There is no meaningful difference — both can be used interchangeably

Q5. An analyst's bar chart has a y-axis that starts at $980,000 instead of $0. A bar representing $1,000,000 appears to be ten times taller than a bar at $982,000. What is the problem?

A) The chart has too many bars to be readable

B) The truncated y-axis exaggerates the visual difference between values

C) The color scheme is inconsistent across bars

D) Bar charts should not be used for dollar values

Q6. Which principle states that effective visualizations should maximize the proportion of ink used to represent actual data vs. decorative elements?

A) The 5-second rule

B) The data-ink ratio

C) Pre-attentive attributes

D) The McCandless Method

Q7. In the McCandless Method, which element represents the specific business outcome the visualization is designed to achieve?

A) Information

B) Story

C) Goal

D) Visual Form

Q8. An analyst builds a Tableau dashboard for a VP of Marketing. According to dashboard design best practices, where should the MOST important KPI be placed?

A) Bottom-right corner for emphasis

B) Center of the dashboard as a focal point

C) Top-left, where the viewer's eye naturally starts

D) In a separate tab to avoid cluttering the main view

Q9. A visualization uses 12 different colors to represent 12 product categories. What is the PRIMARY design problem?

A) Colors must match the company's brand palette

B) Too many colors create confusion and make it impossible to easily distinguish categories

C) Categorical data should always use shades of a single color

D) 12 colors exceed the maximum number allowed in Tableau

Q10. When presenting data findings, which narrative structure is recommended for maximum impact with stakeholders?

A) Data first, then the conclusion at the end

B) Setup (context) → Conflict (problem/tension) → Resolution (insight and recommendation)

C) Methodology first to establish credibility, then data, then conclusion

D) Conclusion first, then all supporting data tables

Q11. In Tableau, how do you change the aggregation of a Measure from SUM to AVG?

A) Delete the field and re-drag it to the view

B) Right-click the pill in the shelf and change the aggregation type

C) Create a new calculated field that divides SUM by COUNT

D) Only SUM aggregation is available for Measures by default

Q12. Which color strategy is MOST appropriate for a chart showing temperature anomalies ranging from -5 degrees (below average) to +5 degrees (above average)?

A) Sequential scale: light to dark shades of blue

B) Categorical colors: a different hue for each value

C) Diverging scale: blue for below average, red for above average, white at zero

D) Monochrome: shades of gray only

CHAPTER 07

Data Analysis with R Programming

R basics · RStudio · tidyverse · dplyr · ggplot2 · R Markdown

14 pages - 12 practice questions

Introduction: R Is Not as Scary as You Think

For most people coming through the GDA certificate, R is the most intimidating part. If you've never written a line of code, a programming language can feel like an impossible wall. But here's the truth: R, as taught in Course 7, is more like learning a new vocabulary than learning to program from scratch. You'll use a small set of well-designed functions that are readable in plain English, and within a few hours of practice, you'll be producing professional visualizations and reports from real data.

R is a statistical programming language used widely in academia, research, and industry for data analysis, statistical modeling, and visualization. The GDA certificate uses R through **RStudio** (an integrated development environment) and focuses on the **tidyverse** — a collection of R packages designed specifically to make data analysis intuitive and expressive.

7.1 R and RStudio Basics

Setting Up

- Download R from **cran.r-project.org** (free)

- Download RStudio Desktop from **posit.co** (free). RStudio is the IDE — it's where you write and run your R code.

- Alternatively: use **Posit Cloud** (formerly RStudio Cloud) — a free browser-based RStudio environment. No installation needed.

The RStudio Interface

Pane	Location	Purpose
Script Editor	Top-left	Write and save R scripts. Run code with Ctrl+Enter (line) or Ctrl+Shift+Enter (script).
Console	Bottom-left	Type and run R commands interactively. Shows output and errors.
Environment / History	Top-right	Shows all variables/objects currently in memory. History shows previous commands.
Files / Plots / Packages / Help	Bottom-right	Browse files, view plots, manage packages, and access documentation.

R Fundamentals

```
# Comments start with #

# Assign values to variables using <- (preferred) or =
x <- 42
name <- "data analyst"
is_certified <- TRUE

# Basic data types
num_value  <- 3.14         # numeric
text_value <- "hello"      # character (string)
bool_value <- TRUE         # logical (boolean)
int_value  <- 5L           # integer

# Print to console
print(x)          # explicit print
x                 # just typing the variable name also prints it

# Vectors: a sequence of values of the same type
scores <- c(85, 92, 78, 95, 88)
regions <- c("West", "East", "North", "South")

# Access elements (R uses 1-based indexing, not 0-based!)
scores[1]        # returns 85 (first element)
scores[2:4]      # returns elements 2 through 4
```

WATCH OUT

R uses 1-based indexing — the first element is [1], not [0] like Python or JavaScript. This is a common source of confusion and off-by-one errors when switching between languages.

7.2 Installing and Loading Packages

```
# Install a package (do this once per machine)
install.packages("tidyverse")

# Load a package (do this at the start of every script)
library(tidyverse)

# The tidyverse loads several packages at once:
# ggplot2 (visualization), dplyr (data manipulation),
# tidyr (data reshaping), readr (reading files), and more.
```

VOCABULARY: Package

A collection of R functions, data, and documentation that extends R's capabilities. Installed once with install.packages(), loaded per session with library().

VOCABULARY: Tidyverse

A collection of R packages designed around a consistent philosophy for data science. Key packages: ggplot2, dplyr, tidyr, readr, purrr, tibble, stringr, forcats.

7.3 Data Manipulation with dplyr

dplyr is the tidyverse package for data manipulation. It provides five core "verbs" that cover most data wrangling tasks:

dplyr Verb	What It Does	SQL Equivalent
filter()	Keep rows that match a condition	WHERE
select()	Keep or drop specific columns	SELECT col1, col2
mutate()	Add new columns or modify existing ones	SELECT *, expression AS new_col
arrange()	Sort rows by column values	ORDER BY
summarize()	Collapse rows into summary statistics	GROUP BY + aggregate
group_by()	Group data for grouped operations (used with summarize)	GROUP BY

The Pipe Operator: %>%

The pipe operator **%>%** (from the magrittr package, included in tidyverse) passes the output of one function as the first input of the next. This lets you chain operations together in a readable left-to-right sequence.

```
library(tidyverse)

# Without pipe (nested, hard to read):
summarize(group_by(filter(sales, region == "West"), category), total = sum(revenue))

# With pipe (same operation, much more readable):
sales %>%
  filter(region == "West") %>%
  group_by(category) %>%
  summarize(total = sum(revenue), avg = mean(revenue)) %>%
  arrange(desc(total))

# Read it as: "Take sales, THEN filter for West, THEN group by category,
# THEN summarize with total and avg, THEN sort by total descending."

# dplyr examples:
# filter: keep only 2025 rows
sales %>% filter(year == 2025)

# select: keep only specific columns
sales %>% select(customer_id, revenue, region)

# mutate: add a profit margin column
sales %>% mutate(margin = (revenue - cost) / revenue * 100)

# arrange: sort by revenue descending
sales %>% arrange(desc(revenue))

# group_by + summarize: revenue by region
sales %>%
  group_by(region) %>%
  summarize(total_revenue = sum(revenue), n_orders = n())
```

PRO TIP

In R 4.1+, there is a native pipe operator |> that works similarly to %>%. The GDA certificate primarily teaches %>%, so use that for the course. They are functionally identical for most use cases.

7.4 Data Visualization with ggplot2

ggplot2 is the tidyverse's visualization package, built on the "Grammar of Graphics" — a systematic framework for describing any chart as a combination of layers. Every ggplot2 chart is built by adding layers with the + operator.

ggplot2 Structure

```
# Basic ggplot2 structure:
ggplot(data = your_dataframe, aes(x = x_column, y = y_column)) +
  geom_type() +
  additional_layers

# aes() = aesthetic mapping: maps data columns to visual properties
# geom_*() = geometric layer: determines the type of chart

# Example: scatter plot
ggplot(data = sales, aes(x = ad_spend, y = revenue)) +
  geom_point()

# Add color mapping by a third variable
ggplot(data = sales, aes(x = ad_spend, y = revenue, color = region)) +
  geom_point(size = 3, alpha = 0.7)

# Line chart: trends over time
ggplot(data = monthly_sales, aes(x = month, y = revenue)) +
  geom_line(color = "#00A693", linewidth = 1) +
  geom_point(size = 2)
```

Common geom Functions

Function	Chart Type	Key Aesthetics (aes)
geom_point()	Scatter plot	x, y, color, size, shape
geom_line()	Line chart	x, y, color, linetype
geom_bar()	Bar chart (counts)	x, fill
geom_col()	Bar chart (values)	x, y, fill
geom_histogram()	Histogram (distribution)	x, fill, bins
geom_boxplot()	Box plot	x, y, fill
geom_smooth()	Trend/regression line	x, y, method

Adding Labels, Titles, and Themes

```
# Complete example: polished scatter plot
ggplot(data = sales, aes(x = ad_spend, y = revenue, color = region)) +
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +  # add linear trend line
  labs(
    title = "Ad Spend vs. Revenue by Region",
    subtitle = "Q1-Q3 2025 | n = 1,247 campaigns",
    x = "Advertising Spend ($)",
    y = "Revenue ($)",
    color = "Region"
  ) +
  theme_minimal() +           # clean, minimal background
  scale_color_brewer(palette = "Set2")  # colorblind-friendly palette
```

VOCABULARY: Grammar of Graphics

The theoretical framework underlying ggplot2. Every visualization is described as a combination of: data, aesthetic mappings (aes), geometric layers (geom), scales, coordinate systems, and themes.

7.5 Data Reshaping with tidyr

tidyr helps you reshape data between wide and long formats — a critical skill when preparing data for visualization with ggplot2.

```
# pivot_longer: wide to long (more rows, fewer columns)
# Use when you have multiple columns that should be one column
sales_long <- sales_wide %>%
  pivot_longer(
    cols = c(Q1, Q2, Q3, Q4),   # columns to convert
    names_to = "quarter",        # new column for old column names
    values_to = "revenue"        # new column for values
  )

# pivot_wider: long to wide (more columns, fewer rows)
# Use when you need to create a cross-tabulation
sales_wide <- sales_long %>%
  pivot_wider(
    names_from = quarter,
    values_from = revenue
  )
```

7.6 R Markdown

R Markdown lets you combine R code, output, and narrative text in a single document that can be exported as HTML, PDF, or Word. It's the standard way to create reproducible, shareable data analysis reports.

~~~
---
title: "Q3 Sales Analysis"
author: "Your Name"
date: "2026-02-15"
output: html_document
---

## Overview

This report analyzes Q3 2025 sales performance across regions.

```{r setup, include=FALSE}
library(tidyverse)
sales <- read_csv("sales_q3_2025.csv")
```

## Revenue by Region

```{r revenue_plot}
sales %>%
  group_by(region) %>%
  summarize(total = sum(revenue)) %>%
  ggplot(aes(x = reorder(region, total), y = total, fill = region)) +
  geom_col() +
  coord_flip() +
  labs(title = "Q3 Revenue by Region", x = "Region", y = "Revenue ($)")
```

Text between code chunks is plain narrative. Knit to HTML with Ctrl+Shift+K.
~~~

VOCABULARY: R Markdown

A file format (.Rmd) that combines R code, code output, and narrative text in one document. Can be "knitted" into HTML, PDF, or Word reports.

VOCABULARY: Reproducible Analysis

An analysis that produces identical results when run again with the same data and code. R Markdown supports reproducibility by keeping code and output together.

WHAT TO REMEMBER

- R is a statistical programming language. RStudio is the IDE. Posit Cloud provides free browser-based access.

- Install packages once with install.packages(). Load per session with library().

- R uses 1-based indexing: the first element is [1], not [0].

- dplyr verbs: filter() = WHERE, select() = SELECT cols, mutate() = add columns, arrange() = ORDER BY, summarize() = aggregation.

- The pipe operator %>% chains operations: read as "take X, THEN do Y, THEN do Z."

- ggplot2: ggplot(data, aes(x, y)) + geom_*(). Every chart is built in layers.

- Common geoms: geom_point() = scatter, geom_line() = line, geom_col() = bar, geom_histogram() = distribution.

- tidyr: pivot_longer() converts wide to long; pivot_wider() converts long to wide.

- R Markdown combines code, output, and narrative. Knit to HTML/PDF/Word with Ctrl+Shift+K.

Chapter 07 — Practice Questions

Answer all 12 questions before checking the Answer Key.

Q1. An analyst wants to keep only rows in a dataframe where the 'region' column equals 'West'. Which dplyr function should they use?

A) select(region == 'West')
B) filter(region == 'West')
C) mutate(region == 'West')

D) arrange(region == 'West')

Q2. What does the pipe operator %>% do in R?

A) It multiplies two numeric values together

B) It passes the output of the left-hand function as the first argument to the right-hand function

C) It concatenates two strings

D) It imports a package into the R session

Q3. An analyst writes: ggplot(data=sales, aes(x=month, y=revenue)) + geom_line(). What type of chart does this create?

A) A bar chart comparing revenue by month

B) A scatter plot of revenue vs. month

C) A line chart showing revenue over months

D) A histogram of the revenue column

Q4. In R, which index retrieves the FIRST element of a vector named scores?

A) scores[0]

B) scores[1]

C) scores.first()

D) scores(1)

Q5. Which dplyr verb is used to ADD a new column to a dataframe?

A) filter()

B) select()

C) mutate()

D) arrange()

Q6. An analyst has a dataframe with columns: customer_id, Q1, Q2, Q3, Q4 (one revenue column per quarter). They want to reshape it so there is one row per customer per quarter. Which tidyr function should they use?

A) pivot_wider()

B) pivot_longer()

C) separate()

D) gather()

Q7. Which of the following correctly installs and then loads the ggplot2 package?

A) install.packages('ggplot2'); import('ggplot2')

B) get.package('ggplot2'); load('ggplot2')

C) install.packages('ggplot2'); library(ggplot2)

D) add.package('ggplot2'); require.package('ggplot2')

Q8. In ggplot2, what is the purpose of the aes() function?

A) It applies an aesthetic theme to the chart (e.g., dark background)

B) It maps data columns to visual properties like x, y, color, and size

C) It sets the font size and color of axis labels

D) It defines the type of geometric layer to use

Q9. An analyst wants to create a bar chart showing total revenue for each region. The data already has a column for total_revenue. Which geom is MOST appropriate?

A) geom_bar() — count-based bars

B) geom_histogram() — distribution

C) geom_col() — value-based bars using a y aesthetic

D) geom_point() — scatter plot

Q10. What is R Markdown used for?

A) Creating interactive Tableau dashboards from R

B) Combining R code, code output, and narrative text in one reproducible document

C) Writing SQL queries that are executed directly in BigQuery

D) Building machine learning models with automated documentation

Q11. An analyst uses the following code: sales %>% group_by(category) %>% summarize(total = sum(revenue), avg = mean(revenue)). What does this return?

A) One row per transaction with the category, revenue sum, and mean

B) One row per unique category with the total and average revenue for that category

C) All rows from sales with two new columns: total and avg

D) A filtered dataset showing only rows where revenue equals the mean

Q12. Which of the following code snippets correctly adds a title, x-axis label, and y-axis label to a ggplot2 chart?

A) ggplot() + title('My Chart') + xlabel('Month') + ylabel('Revenue')

B) ggplot() + set_labels(title='My Chart', x='Month', y='Revenue')

C) ggplot() + labs(title='My Chart', x='Month', y='Revenue')

D) ggplot() + theme(title='My Chart', xlab='Month', ylab='Revenue')

CHAPTER 08

Career Prep + Capstone

Capstone project · portfolio building · resume · LinkedIn · interviews · templates

14 pages - 12 practice questions

Introduction: The Certificate Is the Beginning, Not the End

Completing the Google Data Analytics Certificate is a significant achievement — but by itself, it won't land you a job. What gets you hired is the combination of the certificate, a polished portfolio of real projects, a resume that speaks the language of hiring managers, and the ability to talk about your work confidently in interviews. This chapter covers all of that, starting with the capstone project that wraps up Course 8.

8.1 Completing the Capstone Project

The Course 8 capstone is your chance to demonstrate the full data analytics workflow end-to-end, using a real dataset and a real business scenario. Coursera provides two case study tracks to choose from, and you can also bring your own data for a more personalized project.

The Capstone Structure

- **Ask:** Define the business problem and the questions your analysis will answer.
- **Prepare:** Describe your data sources, assess credibility using ROCCC, and document data limitations.
- **Process:** Document your cleaning steps. List every transformation made and why.
- **Analyze:** Present your analysis — SQL queries, R code, or spreadsheet work — with results.
- **Share:** Create at least 3-5 visualizations that directly support your conclusions.
- **Act:** State 3 clear, actionable recommendations based on your analysis.

PRO TIP

The capstone deliverable is typically a written case study (Google Slides or a PDF) and optionally a Tableau dashboard or R Markdown report. Treat it like a real work deliverable: clear writing, professional visualizations, and specific recommendations. This will be the centerpiece of your portfolio.

8.2 Building Your Data Analytics Portfolio

Your portfolio is what differentiates you from the hundreds of other GDA certificate holders. It shows hiring managers that you can actually do the work — not just that you watched the videos. Aim for **2-4 complete projects** beyond the capstone.

What Makes a Strong Portfolio Project

- **Full workflow:** The best projects show Ask → Prepare → Process → Analyze → Share → Act. Not just a dashboard.
- **Real public datasets:** Use data from Kaggle, Google Dataset Search, or government data portals. Avoid fictional or toy datasets that don't reflect real complexity.
- **A clear business question:** "I analyzed this bike-share dataset to find what factors influence membership conversion" is more compelling than "I explored this dataset."
- **SQL or R code:** Include your actual queries and code. GitHub is the standard place to host this.
- **Visualizations:** At least 3-5 charts that tell a clear story.
- **Written documentation:** A case study document (1-3 pages) explaining your methodology, findings, and recommendations.

Portfolio Platforms

Platform	Best For	Free?
GitHub	Hosting code (R scripts, SQL files), showing technical work	Yes
Kaggle	Publishing notebooks, participating in competitions, visibility	Yes
Tableau Public	Sharing interactive Tableau dashboards publicly	Yes
Personal website	Professional presentation of all projects with written case studies	Low-cost

LinkedIn	Linking to projects, showing certificates and skills	Yes

WATCH OUT

Do not include datasets that contain PII in your public portfolio. Anonymize or use public/synthetic datasets. Also: a Tableau dashboard alone is not a portfolio project. Include the full case study write-up, your data sources, your methodology, and your SQL/R code.

8.3 Writing a Data Analyst Resume

A data analyst resume needs to communicate two things quickly: your technical skills and your impact. Hiring managers spend an average of 7 seconds on initial resume review — make sure the most important information is visible immediately.

Resume Structure for Entry-Level Analysts

- **Header:** Name, email, LinkedIn URL, GitHub/portfolio URL, city/state (no full address needed)
- **Summary (optional):** 2-3 sentences. "Entry-level data analyst with Google Data Analytics Certificate, proficient in SQL, Python, Tableau, and R. 2 portfolio projects analyzing customer behavior and marketing performance."
- **Skills section:** List tools explicitly — SQL, BigQuery, Google Sheets, Tableau, R, Python (if applicable), Excel.
- **Projects:** 2-3 portfolio projects with 2-3 bullet points each. Quantify results where possible.
- **Education/Certifications:** Include Google Data Analytics Professional Certificate with completion date.
- **Work Experience:** Include even non-data roles — focus on any analytical, problem-solving, or data-adjacent tasks.

Writing Strong Bullet Points

Use the **CAR format**: Challenge → Action → Result. Or the **XYZ format**: "Accomplished X as measured by Y by doing Z."

Weak	Strong
Analyzed sales data	Analyzed 18 months of sales data using SQL and Tableau to identify a 23% drop in rep
Created visualizations	Built an interactive Tableau dashboard tracking 8 KPIs for the marketing team, reducing
Used R for analysis	Cleaned and analyzed a 50,000-row customer dataset in R (dplyr/ggplot2), producing fi

8.4 LinkedIn Optimization

- **Headline:** Don't just put "Student" or "Looking for work." Use: "Entry-Level Data Analyst | Google Certified | SQL · Tableau · R"

- **About section:** 3-5 sentences. Who you are, what you do, what you're looking for, and a link to your portfolio.

- **Experience:** Add bullet points quantifying your work even in non-data roles. Did you manage spreadsheets? Analyze performance metrics? Report on KPIs?

- **Skills:** Add SQL, Python, Tableau, R, Google Sheets, Data Analysis, Data Visualization, BigQuery. Request endorsements from connections.

- **Featured section:** Link directly to your GitHub, Tableau Public dashboards, and Kaggle notebooks.

- **Post content:** Share your projects, learnings, and insights. Even 1-2 posts per month dramatically increases profile visibility.

8.5 Interview Preparation

Types of Interview Questions

- **Technical questions:** SQL queries you write or debug on the spot. "Write a query to find the top 5 customers by revenue in each region." Practice on LeetCode, DataLemur, or StrataScratch.

- **Behavioral questions:** "Tell me about a time you had to work with messy data." "Describe a situation where you had to explain a complex finding to a non-technical stakeholder." Use the STAR method.

- **Case questions:** "How would you analyze whether our new feature is increasing user retention?" Think through the six phases of data analysis out loud.

- **Tool-specific questions:** "What's the difference between INNER JOIN and LEFT JOIN?" "When would you use VLOOKUP vs. an INDEX/MATCH?" Review Chapters 4 and 5 before interviews.

The STAR Method for Behavioral Questions

Component	What to Cover
Situation	Set the scene. What was the context? What was the business problem?
Task	What was your specific responsibility in this situation?
Action	What steps did YOU specifically take? Use "I" not "we."
Result	What was the measurable outcome? Quantify where possible.

Common Interview Questions to Prepare For

- "Walk me through your data analytics process from start to finish."

- "How do you handle missing or null values in a dataset?"

- "Tell me about your capstone project. What was your biggest challenge?"

- "What's the difference between a LEFT JOIN and an INNER JOIN?"

- "How would you prioritize competing analysis requests from multiple stakeholders?"

- "What tools are you proficient in and how have you used them in real projects?"
- "Where do you want to be in 2-3 years as a data professional?"

PRO TIP

Practice explaining your portfolio projects out loud — not just having them written down. You should be able to walk an interviewer through any project in 90 seconds: the business problem, the data you used, the analysis you ran, the insight you found, and the recommendation you made. Rehearse this until it's natural.

8.6 Portfolio Project Templates

Use these three templates to structure your portfolio projects. Each template follows the six-phase data analysis framework and includes everything a reviewer or hiring manager would want to see.

TEMPLATE 1: Business Problem Case Study

Section	What to Include
Project Title	A clear, descriptive title: "Analyzing Cyclistic Bike-Share Membership Patterns"
Business Problem	2-3 sentences: What question are you answering? Who is the stakeholder? Why does it matter?
Data Source	Name the dataset, source URL, date range, number of rows, and any limitations.
Tools Used	List: SQL (BigQuery), R (dplyr, ggplot2), Tableau Public, Google Sheets
Data Cleaning	Summarize: What dirty data issues did you find? What did you do to fix them? How many rows removed/modi
Analysis	Describe the key queries/operations you ran. Include 1-2 code snippets.
Key Findings	3-5 bullet points. Each finding should directly answer the business question.
Visualizations	Include 3-5 charts. Each chart needs a title that states the insight, not just the topic.
Recommendations	3 specific, actionable recommendations with your reasoning.
Next Steps	What additional data or analysis would strengthen these findings?

TEMPLATE 2: Data Cleaning Project Write-Up

Section	What to Include

Dataset Overview	Source, size (rows x columns), date range, data types present.
Initial Quality Assessment	What did you find when you first examined the data? Null counts, duplicate counts, format inconsistencies, ou
Cleaning Steps	Numbered list of every cleaning action taken. Include the SQL or R code used for each step.
Changelog	A table: Change | Reason | Before Count | After Count | Date
Validation	How did you verify each cleaning step worked correctly? What checks did you run?
Final Dataset Summary	Row count, null counts, unique value counts in key columns, data types after cleaning.
Reflection	What would you do differently? What data quality issues were hardest to resolve?

TEMPLATE 3: Data Visualization Project Write-Up

Section	What to Include
Project Goal	Who is the audience? What decision should this visualization help them make?
Data Source & Prep	Where did the data come from? What cleaning or transformation was needed before visualization?
Tool Used	Tableau Public / ggplot2 in R / Google Sheets — and why you chose it for this project.
Chart Choices	For each chart: what chart type, why you chose it, what relationship or pattern it shows.
Design Decisions	Color palette, typography, layout — and the reasoning behind each choice.
Key Insights	3-5 findings visible in the visualizations, stated as conclusions not descriptions.
Dashboard Link	Public Tableau link or GitHub link to R Markdown output.
Feedback & Iteration	If you revised any charts based on feedback — what changed and why?

WHAT TO REMEMBER

- The capstone follows all 6 phases: Ask, Prepare, Process, Analyze, Share, Act. Include 3-5 visualizations and clear recommendations.

- Portfolio: 2-4 projects beyond the capstone. Show the full workflow, use public datasets, include code on GitHub.

- Resume: use CAR/XYZ format for bullet points. Quantify results. List all tools explicitly in a skills section.

- LinkedIn: strong headline (not "student"), linked portfolio, skills endorsements, and regular content activity.

- Interview prep: practice SQL queries, STAR stories for behavioral questions, and 90-second project walkthroughs.

- The three portfolio templates cover: business problem case study, data cleaning write-up, and visualization project.

Chapter 08 — Practice Questions

Answer all 12 questions before checking the Answer Key.

Q1. Which of the following BEST describes the purpose of a data analytics portfolio?

A) It is a collection of certificates and credentials from completed courses

B) It demonstrates to hiring managers that you can execute the full data analytics workflow using real data

C) It is a formal academic document required to apply for analyst positions

D) It replaces the need for a resume when applying for entry-level roles

Q2. An analyst is writing their capstone case study. In which section should they include the SQL queries and R code they used during their analysis?

A) Ask phase

B) Prepare phase

C) Process phase

D) Analyze phase

Q3. Which of the following resume bullet points BEST demonstrates strong data analyst experience?

A) 'Worked with large datasets using SQL and visualization tools'

B) 'Passionate about data and committed to continuous learning'

C) 'Analyzed 12 months of e-commerce clickstream data in SQL to identify a 31% cart abandonment rate, resulting in a UX redesign that reduced abandonment by 18%'

D) 'Familiar with SQL, Tableau, and R from Google Data Analytics Certificate'

Q4. What is the STAR method used for in data analytics interviews?

A) Structuring technical SQL answers using Situation, Table, Aggregation, and Result

B) Answering behavioral interview questions with Situation, Task, Action, and Result

C) Evaluating data quality using Structure, Type, Accuracy, and Reliability

D) Building portfolio projects with Story, Tools, Analysis, and Recommendations

Q5. An analyst is preparing for a technical interview and wants to practice SQL. Which platform is MOST appropriate for practicing interview-style SQL questions?

A) Tableau Public

B) GitHub

C) DataLemur or StrataScratch

D) R Markdown

Q6. Which of the following LinkedIn sections has the MOST impact for an entry-level data analyst job search?

A) Education section listing university degrees

B) Connections count (more = better)

C) Headline and Featured section with links to portfolio projects

D) Recommendations from high school teachers

Q7. In the capstone's 'Act' phase, what should an analyst deliver?

A) The raw cleaned dataset ready for stakeholder review

B) A visualization dashboard showing all data collected

C) Specific, actionable recommendations based on the analysis findings

D) A technical appendix documenting every SQL query run

Q8. Which portfolio platform is BEST suited for hosting R scripts and SQL files so that hiring managers can review your code?

A) Tableau Public

B) LinkedIn

C) GitHub

D) Coursera

Q9. An interviewer asks: 'Tell me about a time you had to explain a complex technical finding to a non-technical stakeholder.' Using the STAR method, which component should come FIRST in your answer?

A) Action — describe what you did to simplify the explanation

B) Result — lead with the positive outcome

C) Situation — set the context of the scenario

D) Task — explain your specific role in the project

Q10. According to best practices covered in Chapter 8, how many portfolio projects should an entry-level data analyst aim to have beyond their capstone project?

A) At least 10 to demonstrate breadth of experience

B) 1 is sufficient if it's very detailed

C) 2 to 4 complete projects covering different tools and domains

D) The capstone is sufficient on its own

Q11. A hiring manager reviews two resumes. Candidate A lists 'Proficient in SQL.' Candidate B writes: 'Used SQL (BigQuery) to query 3M+ row datasets, write complex JOINs, and automate weekly reporting queries.' Which candidate demonstrates stronger analytical communication and why?

A) Candidate A — brevity is more professional on a resume

B) Candidate B — specific details and quantified experience demonstrate actual capability

C) Both are equally effective for different types of roles

D) Candidate A — simple language is preferred by ATS systems

Q12. Which of the following is included in the 'Prepare' phase of a capstone write-up?

A) The data visualizations and charts created during the analysis

B) The recommendations made to the business based on analysis findings

C) An assessment of the data source using ROCCC and documentation of data limitations

D) A description of the SQL queries written to analyze the data

Answer Key

All 100 Practice Questions

Answers include a brief explanation to help you understand why each answer is correct — not just what the right letter is.

Chapter 01 — Foundations of Data Analytics

Q1. Answer: B

The data life cycle covers data management (how data is stored, managed, and destroyed). The data analysis process covers what analysts do with data to find insights.

Q2. Answer: B

Data-inspired means using data as one of several inputs — alongside experience and intuition — rather than letting data alone dictate the decision.

Q3. Answer: C

The Process phase is specifically about cleaning data: removing duplicates, handling nulls, fixing errors, and validating the result.

Q4. Answer: C

The VP is an executive stakeholder — someone at a senior level who needs high-level summaries without technical detail.

Q5. Answer: D

Data engineers build and maintain the infrastructure (pipelines, databases) that makes data accessible. DBAs also maintain databases but focus more on performance and security; engineers focus on data movement.

Q6. Answer: C

Creating a logical schema for organizing data is data design — one of the five analytical thinking skills in the GDA certificate.

Q7. Answer: C

BigQuery is Google's SQL-based database platform, used in the GDA certificate for querying large relational datasets.

Q8. Answer: C

The Destroy phase is when data is securely deleted, especially after a retention period expires or when legal requirements mandate deletion.

Q9. Answer: B

Capture is the phase where data is collected or received from sources — including sensors, surveys, and transactions.

Q10. Answer: C

The new analyst is missing understanding of context — the ability to see data within a larger framework (seasonal patterns, industry trends) rather than in isolation.

Chapter 02 — Ask Questions to Make Data-Driven Decisions

Q1. Answer: D

The question is both not specific (what does "performing well" mean?) and not measurable (no criteria for success). SMART questions need both qualities.

Q2. Answer: B

The question specifies a specific behavior (email campaign), a measurable outcome (more spending), and a time frame (60-day period after launch) — making it SMART.

Q3. Answer: C

"Will generate the most leads next quarter" is a forecast — predicting a future outcome. That's a predictive question.

Q4. Answer: B

Structured thinking starts with defining the problem clearly before gathering context, choosing tools, or running analysis.

Q5. Answer: C

Filtering in Google Sheets hides rows that don't match the criteria — they're still in the data but not visible. Nothing is deleted.

Q6. Answer: C

=SUM(B2:B500) is the correct syntax for summing a range in Google Sheets. =COUNT counts cells with numbers; =TOTAL and =ADD do not exist.

Q7. Answer: C

BLUF (Bottom Line Up Front) means leading with the key insight or recommendation, then supporting with data. This is especially important for executive stakeholders.

Q8. Answer: B

The analyst's job is to present findings honestly. Adjusting analysis to match stakeholder preferences is an ethical violation.

Q9. Answer: B

Sorting reorders rows in ascending or descending order. Filtering hides rows that don't meet a condition — the data remains in the sheet.

Q10. Answer: D

Getting everyone to agree on what "improved productivity" means is defining the problem — the first step in structured thinking.

Q11. Answer: C

=COUNTA counts all non-empty cells including text. =COUNT only counts cells with numeric values.

Q12. Answer: C

Option C starts with "Don't you think..." — this is a classic leading question because it pushes toward a specific answer.

Chapter 03 — Prepare Data for Exploration

Q1. Answer: C

Star ratings (1-5) are categorical with a natural order but unequal intervals — that's ordinal data. They're qualitative (descriptive categories with a ranking), not quantitative.

Q2. Answer: B

Only interviewing the most available department is a sampling bias — the sample is not representative of the full employee population.

Q3. Answer: D

Data collected three years ago (before a major regulatory change) fails the "Current" criterion in ROCCC.

Q4. Answer: B

A column in one table that references the primary key of another table is a foreign key. It creates the relationship between tables.

Q5. Answer: C

WHERE is the SQL clause used to filter rows based on a condition. ORDER BY sorts; SELECT specifies columns; LIMIT restricts output count.

Q6. Answer: C

Actively seeking and favoring data that confirms an existing belief is confirmation bias.

Q7. Answer: C

Metadata provides descriptive information about data — who created it, when, what it contains — not the data itself.

Q8. Answer: C

A spreadsheet with rows and columns and a defined structure is structured data. Essays, footage, and audio recordings are unstructured.

Q9. Answer: B

YYYYMMDD format sorts chronologically when files are sorted alphabetically. Files named 20260115 will naturally sort before 20260201, etc.

Q10. Answer: B

The correct syntax for a DESC sort with a LIMIT is: SELECT * FROM table ORDER BY column DESC LIMIT n. Option A is invalid SQL (LIMIT before WHERE).

Q11. Answer: C

Two analysts seeing different patterns in the same dataset is observer bias — different observers interpret the same data differently.

Q12. Answer: B

A data dictionary defines each field in a dataset: name, data type, description, and allowable values.

Chapter 04 — Process Data from Dirty to Clean

Q1. Answer: B

=TRIM() removes leading and trailing spaces, fixing the "Smith " vs "Smith" mismatch. CLEAN removes non-printable characters, which is a separate issue.

Q2. Answer: C

NULL represents the complete absence of a value — it is not zero (a number) or an empty string (a text value with no characters).

Q3. Answer: B

To check for NULL, SQL uses IS NULL or IS NOT NULL. You cannot use = NULL because NULL is not equal to anything, including itself.

Q4. Answer: C

COALESCE returns the first non-NULL value from a list. If the first argument is NULL, it tries the second, then the third, etc.

Q5. Answer: B

COUNTIF counts how many times the value in A2 appears in the range. A result of 3 means the value appears 3 times — indicating duplicates.

Q6. Answer: B

Always create a backup copy before removing duplicates. The operation may remove records that appear to be duplicates but aren't, and it may be irreversible.

Q7. Answer: C

SELECT DISTINCT returns only unique values — duplicates are removed. It does not sort, filter by recency, or count occurrences.

Q8. Answer: C

CAST() converts a value from one data type to another. CAST(column AS INT64) converts to integer.

Q9. Answer: B

Replacing non-standard values ("CA", "Calif.") with a consistent standard ("California") is data standardization — fixing inconsistent data.

Q10. Answer: B

A dropdown list created with Data Validation restricts entry to approved values — preventing incorrect data from being entered.

Q11. Answer: C

Checking row count before and after cleaning to confirm the expected number of rows were affected is post-cleaning validation.

Q12. Answer: B

A complete changelog entry includes: date, description of change, reason for change, who made it, and before/after counts to document the impact.

Q13. Answer: C

=CLEAN() removes non-printable characters imported from legacy systems. =TRIM() removes whitespace, which is a different problem.

Q14. Answer: B

Data integrity is the accuracy, completeness, consistency, and trustworthiness of data throughout its entire life cycle.

Chapter 05 — Analyze Data to Answer Questions

Q1. Answer: C

SELECT region, COUNT(*) FROM orders GROUP BY region returns one row per region with the count of orders in that region.

Q2. Answer: C

HAVING filters grouped results after aggregation. WHERE cannot use aggregate functions. HAVING COUNT(*) > 500 correctly filters after grouping.

Q3. Answer: B

LEFT JOIN keeps all rows from the left table. Where there's no match in the right table (Orders), the order columns contain NULL.

Q4. Answer: B

WHERE orders.order_id IS NULL identifies customers whose join to Orders produced no match — meaning they have never placed an order.

Q5. Answer: D

FULL OUTER JOIN returns all rows from both tables, with NULLs filling in the missing side for any unmatched rows.

Q6. Answer: C

Each cell in a pivot table with SUM aggregation shows the total revenue for the specific row-column intersection (that category in that quarter).

Q7. Answer: B

The range starts at column A (customer_id). Column B (name) is index 2. Use FALSE for exact match. =VLOOKUP(A2, Sheet2!A:C, 2, FALSE).

Q8. Answer: B

SUMIF(B2:B500, "West", C2:C500) sums the values in C2:C500 only for rows where the corresponding B column value equals "West".

Q9. Answer: B

SQL always executes the subquery first, then passes its result to the outer query. The inner query must complete before the outer can run.

Q10. Answer: C

To compare each row against the overall average, the WHERE clause must contain a subquery: WHERE total > (SELECT AVG(total) FROM orders). You cannot use AVG() directly in WHERE.

Q11. Answer: B

GROUP BY already returns unique category values — adding DISTINCT is redundant and unnecessary. Note: the query would also need an aggregate function to be valid.

Q12. Answer: B

The Filters area restricts which data is included in the entire pivot table calculation, based on the values of the filtered field.

Q13. Answer: B

SQL order of operations: WHERE filters rows first, then GROUP BY groups them, then HAVING filters groups, then ORDER BY sorts the final output.

Q14. Answer: C

1,800 matching pairs were returned. Option D is also partially correct (B is also true: 200 orders have no matching customer), making C the most complete answer.

Q15. Answer: B

COUNTIFS (plural) accepts multiple range-criteria pairs. COUNTIF (singular) only accepts one condition. The syntax is COUNTIFS(range1, criteria1, range2, criteria2).

Q16. Answer: B

A subquery in the SELECT clause is a scalar subquery — it returns a single value (like the overall average) that is applied to every row in the result.

Chapter 06 — Share Data Through the Art of Visualization

Q1. Answer: C

Line charts are specifically designed to show trends over time. They're the standard choice when x-axis = time periods.

Q2. Answer: B

Pie charts with more than 5 segments are very difficult to read because human perception cannot accurately distinguish many similarly-sized slices.

Q3. Answer: C

Scatter plots show the relationship (correlation) between two continuous variables. They're ideal for 200 campaigns with two numeric values each.

Q4. Answer: B

In Tableau: Dimensions are categorical (blue), Measures are numeric (green) and can be aggregated with SUM, AVG, COUNT, etc.

Q5. Answer: B

Truncating the y-axis (starting at $980K instead of $0) makes small differences appear much larger, misleading the viewer about the actual magnitude of difference.

Q6. Answer: B

The data-ink ratio principle (Edward Tufte) states you should maximize the proportion of ink used to represent data vs. decorative elements.

Q7. Answer: C

In McCandless Method, Goal = the specific outcome the visualization serves — the decision it helps make or the business purpose it fulfills.

Q8. Answer: C

Standard visual reading pattern (F-pattern/Z-pattern) means viewers naturally start at the top-left. The most important metrics should anchor there.

Q9. Answer: B

Using 12 different colors overloads the viewer's working memory and makes it nearly impossible to track which category corresponds to which color.

Q10. Answer: B

The Setup-Conflict-Resolution narrative structure (from data storytelling) maximizes impact by creating narrative tension before the resolution.

Q11. Answer: B

Right-clicking the measure pill in the shelf shows aggregation options including SUM, AVG, COUNT, MIN, MAX, and others.

Q12. Answer: C

Diverging scales (two contrasting colors with a neutral midpoint) are ideal for data with a meaningful center point, like anomalies above and below average.

Chapter 07 — Data Analysis with R Programming

Q1. Answer: B

filter() keeps rows that match a condition — it's the dplyr equivalent of SQL's WHERE clause.

Q2. Answer: B

The pipe operator %>% takes the output on its left and passes it as the first argument to the function on its right — enabling readable left-to-right chaining.

Q3. Answer: C

geom_line() creates a line chart. With x=month and y=revenue, this plots revenue over months as a connected line.

Q4. Answer: B

R uses 1-based indexing. scores[1] returns the first element. scores[0] returns an empty vector (not an error, but not useful).

Q5. Answer: C

mutate() adds new columns to a dataframe. select() chooses columns; filter() filters rows; arrange() sorts rows.

Q6. Answer: B

pivot_longer() converts wide format (multiple columns representing one variable) to long format (one column for the variable, one for its values).

Q7. Answer: C

install.packages("ggplot2") installs it once; library(ggplot2) loads it for the current session. These are the correct R functions.

Q8. Answer: B

aes() (aesthetic mapping) maps columns from the data to visual properties in the chart: x-axis, y-axis, color, size, shape, etc.

Q9. Answer: C

geom_col() creates bars using a y aesthetic for the bar height. geom_bar() counts rows automatically and doesn't take a y aesthetic.

Q10. Answer: B

R Markdown combines code, output, and narrative in one reproducible document that can be exported as HTML, PDF, or Word.

Q11. Answer: B

group_by(category) groups the data by category, then summarize() collapses each group into one row with total revenue and average revenue.

Q12. Answer: C

labs() is the ggplot2 function for adding labels: title, subtitle, x, y, color, etc. The other options use incorrect syntax.

Chapter 08 — Career Prep + Capstone

Q1. Answer: B

A portfolio demonstrates practical ability — that you can execute the full analytical workflow on real data. Certificates show completion; portfolios show capability.

Q2. Answer: D

The Analyze phase is where you execute your analysis — running queries, writing code, and producing results. The Process phase is for cleaning and transforming data.

Q3. Answer: C

Option C uses the XYZ/CAR format, includes specific tools, quantifies the work (12 months, 3M rows), states a concrete finding (31%), and shows business impact (18% reduction).

Q4. Answer: B

STAR = Situation, Task, Action, Result. It's the standard framework for answering behavioral interview questions with a structured, complete story.

Q5. Answer: C

DataLemur and StrataScratch provide real interview SQL questions from companies like Google, Meta, and Amazon — ideal for technical interview preparation.

Q6. Answer: C

The LinkedIn headline is the first thing recruiters see in search results. The Featured section lets you directly link to portfolio work, making it immediately accessible.

Q7. Answer: C

The Act phase is where analysis turns into action: specific, concrete recommendations that stakeholders can act on based on the findings.

Q8. Answer: C

GitHub is the standard platform for hosting code files (R scripts, SQL queries, Python notebooks) so hiring managers and technical reviewers can evaluate your work.

Q9. Answer: C

STAR starts with Situation — you need to establish context before explaining your task, actions, and results. Context makes the rest of the story meaningful.

Q10. Answer: C

2-4 projects is the recommended range — enough to show breadth and depth without being so many that quality suffers. Each should cover different tools and domains.

Q11. Answer: B

Specific details and quantified experience ("3M+ row datasets," "complex JOINs," "automate weekly reporting") demonstrate actual analytical capability, not just familiarity.

Q12. Answer: C

The Prepare phase covers data sourcing and assessment — describing where the data came from, assessing credibility with ROCCC, and documenting any limitations.

SQL Quick Reference

Every SQL concept tested in the Google Data Analytics Certificate

This condensed reference covers all SQL syntax used in the GDA certificate. Use it during study and keep it handy in the week before your assessments.

Query Structure & Clause Order

SQL clauses must be written in this order (though not all are required in every query):

```
SELECT   column1, column2, aggregate_function(column3)
FROM     table_name
JOIN     other_table ON table_name.id = other_table.id
WHERE    condition                   -- filters rows BEFORE grouping
GROUP BY column1, column2
HAVING   aggregate_condition         -- filters groups AFTER aggregation
ORDER BY column1 ASC|DESC
LIMIT    n;
```

SELECT, FROM, WHERE

```
-- Select all columns
SELECT * FROM orders;

-- Select specific columns with aliases
SELECT customer_id AS id,
       order_date  AS date,
       total_amount AS revenue
FROM orders;

-- Filter rows (WHERE)
SELECT * FROM orders
WHERE total_amount > 100
  AND region = 'West'
  AND order_date >= '2025-01-01';

-- Pattern matching with LIKE
SELECT * FROM customers
WHERE email LIKE '%@gmail.com';   -- ends with @gmail.com
WHERE name  LIKE 'J%';            -- starts with J
WHERE name  LIKE '%son%';         -- contains 'son'

-- Match a list of values with IN
SELECT * FROM orders
WHERE region IN ('West', 'South', 'Central');

-- Exclude a list with NOT IN
SELECT * FROM orders
WHERE status NOT IN ('Cancelled', 'Returned');

-- Range filter with BETWEEN (inclusive)
SELECT * FROM orders
WHERE total_amount BETWEEN 50 AND 200;

-- NULL checks (never use = NULL)
SELECT * FROM customers WHERE phone IS NULL;
SELECT * FROM customers WHERE phone IS NOT NULL;
```

Aggregation & GROUP BY / HAVING

```
-- Aggregate functions
SELECT
  COUNT(*)                AS total_rows,
  COUNT(DISTINCT cust_id) AS unique_customers,
  SUM(revenue)            AS total_revenue,
  AVG(revenue)            AS avg_revenue,
  MIN(revenue)            AS min_revenue,
  MAX(revenue)            AS max_revenue
FROM orders;

-- GROUP BY: aggregate per category
SELECT region, COUNT(*) AS orders, SUM(revenue) AS total
FROM orders
GROUP BY region
ORDER BY total DESC;

-- HAVING: filter groups (use after GROUP BY)
SELECT region, SUM(revenue) AS total
FROM orders
GROUP BY region
HAVING SUM(revenue) > 100000;

-- Combine WHERE (row filter) + HAVING (group filter)
SELECT region, COUNT(*) AS cnt, SUM(revenue) AS total
FROM orders
WHERE year = 2025
GROUP BY region
HAVING COUNT(*) > 50
ORDER BY total DESC;
```

JOINs

JOIN Type	Returns	NULL Filled From
INNER JOIN	Only rows matching in BOTH tables	Neither — unmatched excluded
LEFT JOIN	All left rows + matching right rows	Right table
RIGHT JOIN	All right rows + matching left rows	Left table
FULL OUTER JOIN	All rows from both tables	Both sides
CROSS JOIN	Every left row paired with every right row	N/A — no join condition

```
-- INNER JOIN: only matching rows
SELECT c.name, o.order_id, o.total
FROM customers c
INNER JOIN orders o ON c.customer_id = o.customer_id;

-- LEFT JOIN: all customers, even those with no orders
SELECT c.name, o.order_id
FROM customers c
LEFT JOIN orders o ON c.customer_id = o.customer_id;

-- Find customers with NO orders (unmatched left rows)
SELECT c.name
FROM customers c
LEFT JOIN orders o ON c.customer_id = o.customer_id
WHERE o.order_id IS NULL;

-- Join three tables
SELECT c.name, o.order_id, p.product_name
FROM customers c
INNER JOIN orders o    ON c.customer_id  = o.customer_id
INNER JOIN products p  ON o.product_id   = p.product_id;
```

Subqueries

```
-- Subquery in WHERE: rows above average
SELECT customer_id, total_spent
FROM customers
WHERE total_spent > (SELECT AVG(total_spent) FROM customers);

-- Subquery in FROM (inline view / derived table)
SELECT region, AVG(order_total) AS avg_order
FROM (
    SELECT region, order_total
    FROM orders
    WHERE year = 2025
) AS recent_orders
GROUP BY region;

-- Correlated subquery: references outer query per row
SELECT order_id, total,
  (SELECT AVG(total) FROM orders) AS overall_avg
FROM orders;

-- EXISTS: returns rows where subquery finds at least one match
SELECT customer_id
FROM customers c
WHERE EXISTS (
    SELECT 1 FROM orders o
    WHERE o.customer_id = c.customer_id
      AND o.total > 500
);
```

Cleaning Functions

```
-- Remove whitespace
SELECT TRIM(customer_name)           AS clean_name FROM customers;
SELECT LTRIM(RTRIM(customer_name))   AS clean_name FROM customers;

-- NULL handling
SELECT COALESCE(phone, 'Unknown')    AS phone FROM customers;
SELECT IFNULL(discount, 0)           AS discount FROM orders;

-- Type conversion
SELECT CAST(price_text AS FLOAT64)   AS price FROM products;
SELECT CAST(order_date  AS DATE)      AS date  FROM orders;

-- Conditional transformation
SELECT order_id,
  CASE
    WHEN status = 'C'  THEN 'Completed'
    WHEN status = 'P'  THEN 'Pending'
    WHEN status = 'X'  THEN 'Cancelled'
    ELSE 'Unknown'
  END AS status_label
FROM orders;

-- Deduplication
SELECT DISTINCT customer_id FROM orders;

-- String functions (BigQuery)
SELECT UPPER(name), LOWER(name), LENGTH(name) FROM customers;
SELECT SUBSTR(name, 1, 3) FROM customers;  -- first 3 chars
```

Date Functions (BigQuery)

```
-- Current date / time
SELECT CURRENT_DATE(), CURRENT_TIMESTAMP();

-- Extract parts from a date
SELECT
  EXTRACT(YEAR  FROM order_date) AS yr,
  EXTRACT(MONTH FROM order_date) AS mo,
  EXTRACT(DAY   FROM order_date) AS dy
FROM orders;

-- Date arithmetic
SELECT DATE_ADD(order_date, INTERVAL 30 DAY) AS due_date FROM orders;
SELECT DATE_DIFF(CURRENT_DATE(), order_date, DAY) AS days_ago FROM orders;

-- Format a date
SELECT FORMAT_DATE('%Y-%m', order_date) AS month FROM orders;
```

Window Functions (Advanced — Good to Know)

```
-- ROW_NUMBER: sequential rank (no ties)
SELECT customer_id, revenue,
  ROW_NUMBER() OVER (PARTITION BY region ORDER BY revenue DESC) AS rank
FROM orders;

-- RANK: rank with gaps for ties
-- DENSE_RANK: rank without gaps for ties

-- Running total
SELECT order_date, revenue,
  SUM(revenue) OVER (ORDER BY order_date) AS running_total
FROM orders;

-- LAG/LEAD: compare to previous/next row
SELECT order_date, revenue,
  LAG(revenue, 1) OVER (ORDER BY order_date) AS prev_revenue,
  revenue - LAG(revenue, 1) OVER (ORDER BY order_date) AS change
FROM monthly_sales;
```

Spreadsheet Formula Cheat Sheet

Essential Google Sheets formulas for the GDA certificate

Lookup & Reference Functions

Formula	Syntax	What It Does
VLOOKUP	=VLOOKUP(key, range, col_index, FALSE)	Looks up a value in the first column of a range, returns value from sp
HLOOKUP	=HLOOKUP(key, range, row_index, FALSE)	Same as VLOOKUP but searches horizontally across the first row.
INDEX	=INDEX(range, row_num, col_num)	Returns the value at a specific row/column intersection within a rang
MATCH	=MATCH(value, range, 0)	Returns the position of a value within a range. Use 0 for exact match
INDEX+MATCH	=INDEX(return_range, MATCH(key, lookup_range, 0))	More flexible than VLOOKUP — can look left, handles column insert
XLOOKUP	=XLOOKUP(key, lookup_arr, return_arr)	Modern replacement for VLOOKUP. Returns full rows, handles not-f

```
-- VLOOKUP: return customer name from a lookup table
=VLOOKUP(A2, Sheet2!$A:$C, 2, FALSE)
-- A2 = search key | Sheet2!A:C = range | 2 = return col 2 | FALSE = exact

-- INDEX+MATCH: more flexible lookup (can look left)
=INDEX(Sheet2!$B:$B, MATCH(A2, Sheet2!$A:$A, 0))
-- Returns col B value where col A matches A2
```

WATCH OUT

VLOOKUP index counts from the START of your range — not from column A. If your range is D:G and you want column F, the index is 3 (not 6). Always use FALSE (exact match) unless working with sorted data for approximate matching.

Conditional Functions

Formula	Syntax	Example
IF	=IF(condition, true_val, false_val)	=IF(B2>100, "High", "Low")
IFS	=IFS(cond1, val1, cond2, val2, ...)	=IFS(B2>100,"High",B2>50,"Mid",TRUE,"Low")
COUNTIF	=COUNTIF(range, criteria)	=COUNTIF(B2:B500, "West")
COUNTIFS	=COUNTIFS(r1, c1, r2, c2, ...)	=COUNTIFS(B2:B500,"West",C2:C500,">100")
SUMIF	=SUMIF(range, criteria, sum_range)	=SUMIF(B2:B500, "West", C2:C500)
SUMIFS	=SUMIFS(sum_r, r1, c1, r2, c2, ...)	=SUMIFS(C2:C500,B2:B500,"West",D2:D500,">0")
AVERAGEIF	=AVERAGEIF(range, criteria, avg_range)	=AVERAGEIF(B2:B500, "West", C2:C500)

Statistical & Math Functions

Formula	What It Does	Example
=SUM(range)	Adds all numeric values	=SUM(B2:B500)
=AVERAGE(range)	Calculates the mean	=AVERAGE(C2:C100)
=MEDIAN(range)	Returns the middle value	=MEDIAN(D2:D200)
=COUNT(range)	Counts cells with numbers	=COUNT(A2:A500)
=COUNTA(range)	Counts non-empty cells (including text)	=COUNTA(A2:A500)
=COUNTBLANK(range)	Counts empty cells	=COUNTBLANK(A2:A500)
=MAX(range)	Returns the largest value	=MAX(E2:E50)
=MIN(range)	Returns the smallest value	=MIN(E2:E50)
=STDEV(range)	Standard deviation of a sample	=STDEV(C2:C100)
=ROUND(num, digits)	Rounds to specified decimal places	=ROUND(C2, 2)
=ABS(num)	Returns the absolute value	=ABS(B2-C2)

Text Functions

Formula	What It Does	Example
=TRIM(text)	Removes leading/trailing spaces	=TRIM(A2)

=CLEAN(text)	Removes non-printable characters	=CLEAN(A2)
=UPPER(text)	Converts to ALL CAPS	=UPPER(A2)
=LOWER(text)	Converts to lowercase	=LOWER(A2)
=PROPER(text)	Capitalizes First Letter Of Each Word	=PROPER(A2)
=LEN(text)	Returns number of characters	=LEN(A2)
=LEFT(text, n)	Returns first n characters	=LEFT(A2, 3)
=RIGHT(text, n)	Returns last n characters	=RIGHT(A2, 4)
=MID(text, start, n)	Returns n characters from position start	=MID(A2, 3, 5)
=CONCATENATE()	Joins text strings together	=CONCATENATE(A2," ",B2)
=TEXT(value, format)	Formats a number as text	=TEXT(A2,"$#,##0.00")
=SUBSTITUTE()	Replaces text within a string	=SUBSTITUTE(A2,"CA","California")

Pivot Table Quick Reference

Step	Action	Notes
1. Select data	Click any cell in your dataset	Include the header row; no blank columns
2. Insert pivot	Insert > Pivot table	Choose New sheet for a clean view
3. Add Rows	Drag a dimension to the Rows area	Creates one row per unique value
4. Add Values	Drag a measure to the Values area	Default = SUM; right-click to change
5. Add Columns	Drag a second dimension to Columns	Creates a cross-tabulation matrix
6. Add Filters	Drag a field to the Filters area	Adds a dropdown to filter all pivot data
7. Sort	Click the Row/Column dropdown arrow	Sort by value ascending or descending
8. Show %	Right-click a value cell > Show as	"% of grand total" shows proportions

PRO TIP

The most common pivot table mistake: forgetting to change the aggregation from SUM to COUNTA when you're counting records (e.g., "how many orders per region?"). SUM of order IDs returns meaningless totals; COUNTA counts how many rows there are.

ANNEX

Additional Resources & Deep Reference Material

This annex goes beyond the 8 courses to give you everything you need to pass with confidence, launch your career, and keep growing after the certificate.

A1 — Comprehensive Glossary

This A–Z glossary covers every key term used across all 8 courses of the GDA certificate. Definitions are written in plain English — the way you'd explain them to a friend, not copied from a textbook.

Aggregate Function

A SQL or spreadsheet function that performs a calculation on multiple rows and returns a single result. Examples: COUNT, SUM, AVG, MIN, MAX.

Anonymization

The process of removing or masking personally identifying information from a dataset so individuals cannot be identified.

Attribute

A characteristic or quality of data. In a database table, each column is an attribute. Example: customer_name, order_date.

Bias

A systematic error that causes data to be unrepresentative or analysis to be skewed. Types include sampling bias, confirmation bias, and observer bias.

BigQuery

Google's cloud-based data warehouse and SQL query engine, used in the GDA certificate for running SQL queries on large datasets. Free sandbox access available.

Boolean

A data type with only two possible values: TRUE or FALSE. Used for yes/no, on/off, and binary conditions.

CAST

A SQL function that converts a value from one data type to another. Example: CAST(price_text AS FLOAT64).

Changelog

A document that records every change made to a dataset, including what was changed, when, by whom, and why.

Cleaned Data

Data that has been processed to remove errors, duplicates, null values, and inconsistencies so it is ready for analysis.

COALESCE

A SQL function that returns the first non-NULL value from a list of arguments. Used to replace NULLs with default values.

Conditional Formatting

A Google Sheets feature that automatically applies formatting (color, bold, etc.) to cells that meet a specified condition.

Confirmation Bias

The tendency to search for, favor, or interpret information in a way that confirms one's existing beliefs or hypotheses.

Continuous Data

Numeric data that can take any value within a range, including decimals. Example: temperature, weight, time elapsed.

COUNTIF

A Google Sheets formula that counts cells in a range that meet a specific condition. Syntax: =COUNTIF(range, criteria).

Dashboard

A visual display of multiple charts and KPIs on a single screen, giving stakeholders an at-a-glance view of key metrics.

Data Analyst

A professional who collects, cleans, analyzes, and visualizes data to answer business questions and support decision-making.

Data Dictionary

A document that defines each field in a dataset — field name, data type, description, and allowable values.

Data-Driven Decision-Making

Using facts, metrics, and evidence from data analysis to directly guide business strategy and decisions.

Data Engineer

A professional who designs and maintains the data infrastructure — pipelines, warehouses, and databases — that makes data available for analysis.

Data Ethics

The principles and practices for collecting, storing, and using data in ways that are fair, legal, transparent, and respect individual privacy.

Data Integrity

The accuracy, completeness, consistency, and trustworthiness of data throughout its entire life cycle.

Data Life Cycle

The stages data goes through from creation to destruction: Plan, Capture, Manage, Analyze, Archive, Destroy.

Data Science

A broader field that combines statistics, programming, and domain expertise to extract insights and build predictive models from data.

Data Type

The category of a value in a dataset: numeric, string/text, boolean, date, etc. Determines how the data can be stored and analyzed.

Data Visualization

The graphical representation of data using charts, graphs, and maps to communicate patterns and insights clearly.

Descriptive Analytics

Analysis that answers "what happened?" — summarizing historical data to understand past performance.

Diagnostic Analytics

Analysis that answers "why did it happen?" — identifying causes and patterns behind observed outcomes.

Dimension

In Tableau, a categorical field (shown in blue) used to group, segment, or label data. Examples: Region, Product Category.

Discrete Data

Countable numeric data that can only take specific values, typically whole numbers. Example: number of orders, headcount.

DISTINCT

A SQL keyword that removes duplicate rows from query results. SELECT DISTINCT returns only unique values.

dplyr

An R package in the tidyverse for data manipulation. Key verbs: filter(), select(), mutate(), arrange(), summarize(), group_by().

Duplicate

A record in a dataset that appears more than once. Duplicates must be identified and addressed during data cleaning.

ETL

Extract, Transform, Load — the process of pulling data from a source system, transforming it for analysis, and loading it into a destination.

Filter

In spreadsheets: hides rows that don't match a condition (data remains in the sheet). In SQL: the WHERE or HAVING clause.

Foreign Key (FK)

A column in one table that references the primary key of another table, creating a relationship between them.

FULL OUTER JOIN

A SQL join that returns all rows from both tables, with NULLs filling in the missing side for any unmatched rows.

ggplot2

An R visualization package built on the Grammar of Graphics. Charts are built by adding layers: ggplot(data, aes()) + geom_*().

Grammar of Graphics

The theoretical framework underlying ggplot2 — any chart can be described as data + aesthetic mappings + geometric layers + scales.

GROUP BY

A SQL clause that groups rows sharing a common value so aggregate functions can be applied to each group.

HAVING

A SQL clause that filters grouped results after aggregation. Used after GROUP BY. Contrast with WHERE (filters before grouping).

Inner Join

A SQL join that returns only rows with matching values in both tables. Non-matching rows from either table are excluded.

INNER JOIN

See Inner Join.

KPI

Key Performance Indicator — a measurable value that demonstrates how effectively an organization is achieving a key objective.

Left Join

A SQL join that returns all rows from the left table plus matching rows from the right. Where no match exists, right-side columns are NULL.

LIMIT

A SQL clause that restricts the number of rows returned. Example: LIMIT 10 returns only the first 10 rows.

Measure

In Tableau, a numeric field (shown in green) that can be aggregated. Examples: Revenue, Order Count, Profit.

Metadata

Data about data. Describes a dataset's content, structure, origin, format, and management properties.

Nominal Data

Categorical data with no natural order or ranking. Example: product color (red, blue, green), country name.

NULL

A special marker in a database representing a missing or unknown value. Not the same as zero or an empty string.

Observer Bias

When different people interpreting the same data reach different conclusions based on their own perspectives and expectations.

Ordinal Data

Categorical data with a natural order or ranking, but with unequal intervals between values. Example: satisfaction ratings 1–5.

PII

Personally Identifiable Information — any data that can be used to identify a specific individual. Must be protected and handled ethically.

Pivot Table

A spreadsheet tool that cross-tabulates and summarizes data interactively, grouping rows, columns, and aggregating values.

Pre-attentive Attributes

Visual properties processed by the brain automatically before conscious thought: color, size, shape, position, length.

Primary Key (PK)

A column (or set of columns) that uniquely identifies each row in a database table. No two rows can share a primary key.

Qualitative Data

Descriptive, non-numeric data about qualities or characteristics. Example: customer feedback text, product category.

Quantitative Data

Numeric data that can be counted or measured mathematically. Example: revenue, temperature, order count.

Query

A request for data from a database, written in SQL. A SELECT query retrieves data; other queries can insert, update, or delete.

R

A statistical programming language used for data analysis, visualization, and modeling. Taught in Course 7 of the GDA certificate.

R Markdown

A file format (.Rmd) that combines R code, output, and narrative text in one document, exportable as HTML, PDF, or Word.

Relational Database

A database that organizes data into tables linked by shared key fields (primary keys and foreign keys).

ROCCC

A framework for evaluating data credibility: Reliable, Original, Comprehensive, Current, Cited.

RStudio

An integrated development environment (IDE) for R programming. Free desktop and browser-based (Posit Cloud) versions available.

Sampling Bias

When a data sample is not representative of the full population, leading to skewed or invalid conclusions.

Schema

The structural blueprint of a database — which tables exist, what columns they have, their data types, and how they relate.

SMART Questions

Questions that are Specific, Measurable, Action-oriented, Relevant, and Time-bound. Framework for guiding effective data analysis.

SQL

Structured Query Language — the standard language for querying and managing relational databases.

Structured Data

Data organized in defined rows and columns with a consistent schema. Easily stored in spreadsheets and SQL databases.

Subquery

A query nested inside another query. The inner query runs first and its result is used by the outer query.

SUMIF

A Google Sheets formula that sums values in a range where a corresponding range meets a condition. =SUMIF(range, criteria, sum_range).

Tableau

A data visualization platform used in Course 6. Tableau Public is the free version used in the GDA certificate.

Tidyverse

A collection of R packages (ggplot2, dplyr, tidyr, etc.) designed around a consistent philosophy for data science.

TRIM

Removes leading and trailing whitespace from text. Available in both Google Sheets (=TRIM()) and SQL (TRIM()).

Unstructured Data

Data with no predefined format: images, audio, video, email body text, social media posts.

VLOOKUP

A Google Sheets function that searches the first column of a range for a key and returns a value from a specified column. =VLOOKUP(key, range, index, FALSE).

WHERE

A SQL clause that filters rows based on a condition, applied before any grouping. Contrast with HAVING.

A2 — Common Exam Mistakes & How to Avoid Them

Compiled from Reddit (r/GoogleCertificates, r/dataanalysis), Coursera discussion boards, and student reviews — these are the mistakes that trip up students most often. Read these carefully before your assessments.

Confusing WHERE and HAVING

WHERE filters rows before grouping. HAVING filters groups after aggregation. You cannot use aggregate functions (like COUNT(*) or SUM()) in a WHERE clause — that always goes in HAVING. Memory trick: WHERE comes before GROUP BY in your SQL; HAVING comes after.

Using = NULL instead of IS NULL

In SQL, NULL is not a value — it's the absence of a value. NULL is not equal to anything, not even itself. So WHERE column = NULL never returns rows. Always use WHERE column IS NULL or WHERE column IS NOT NULL.

Getting the VLOOKUP index wrong

The index number in VLOOKUP counts from the START of your selected range — not from column A. If your range is D:G and you want column F, the index is 3, not 6. Draw a small diagram counting columns in your range if you're unsure.

Confusing the data life cycle with the data analysis process

Life cycle = how data is managed (Plan, Capture, Manage, Analyze, Archive, Destroy). Analysis process = what analysts do (Ask, Prepare, Process, Analyze, Share, Act). Quizzes describe scenarios and ask which phase applies — know both frameworks cold.

Thinking LEFT JOIN = "left table only"

A LEFT JOIN returns ALL rows from the left table PLUS matching rows from the right table. Unmatched right-side columns are NULL. It is NOT the same as SELECT * FROM left_table. The most common mistake is forgetting that unmatched left rows still appear — with NULLs in the right-side columns.

Mixing up nominal and ordinal data

Nominal = categories with no natural order (colors, countries, product types). Ordinal = categories WITH a natural order (star ratings, survey scales, education levels). The key question: can you meaningfully say one category is "greater than" another?

Sampling bias vs. confirmation bias

Sampling bias is about who is in your dataset (non-representative group). Confirmation bias is about analyst behavior (only looking for supporting evidence). Quiz scenarios describe a situation — identify whether the problem is in data collection (sampling) or in how the analyst is using the data (confirmation).

Forgetting to use FALSE in VLOOKUP

VLOOKUP with TRUE (or omitting the last argument) does approximate matching — it assumes your first column is sorted and finds the nearest match. For data analysis you almost always want exact matches. Always specify FALSE.

Pivot table counts instead of sums (or vice versa)

The Values area in a pivot table defaults to SUM for numeric columns. If you drag a text or ID column to Values, it defaults to COUNTA. If your pivot table numbers look wrong, check the aggregation type by clicking the Values field and changing the summarization.

R indexing starting at 0 instead of 1

Unlike Python and JavaScript, R uses 1-based indexing. The first element of a vector is scores[1], not scores[0]. Accessing scores[0] in R returns an empty vector — it does not return the first element.

Forgetting GROUP BY rules in SQL

Every column in your SELECT that is NOT inside an aggregate function must appear in the GROUP BY clause. Selecting customer_name and region but only grouping by customer_name will cause an error.

Misidentifying the "Act" phase of analysis

Many students think "Act" means implementing something technically. In the GDA certificate, Act means delivering your recommendations to stakeholders so they can make an informed decision. The analyst's deliverable in the Act phase is insights and recommendations — not code or technical implementation.

A3 — Data Analytics Job Market Overview

The data analytics job market is one of the strongest in tech — and the GDA certificate is increasingly recognized by employers as a meaningful credential. Here's what the market looks like for certificate graduates in 2026.

Salary Expectations

Role / Level	Typical Salary Range (US)	Notes
Junior / Entry-Level Data Analyst	$55,000 – $80,000	GDA cert most relevant here
Mid-Level Data Analyst (2–4 yrs)	$75,000 – $105,000	Python + SQL proficiency valued
Senior Data Analyst (5+ yrs)	$100,000 – $140,000	Domain expertise + leadership
Business Intelligence Analyst	$70,000 – $110,000	Dashboard + reporting focus
Data Scientist (entry-level)	$90,000 – $130,000	Requires ML/statistics background
Analytics Engineer	$95,000 – $145,000	dbt, SQL pipelines, data modeling

Location matters enormously. Salaries in San Francisco, New York, and Seattle run 40–60% above national averages. Remote roles have normalized pay bands somewhat, but top-paying companies still skew toward tech hubs. Always research the specific company and location using Glassdoor, Levels.fyi, or LinkedIn Salary.

Job Titles to Search For

Entry-level analysts are often listed under many different titles. Search for all of these:

- **Data Analyst** — the most common title
- **Junior Data Analyst / Associate Data Analyst** — explicitly entry-level
- **Business Analyst** — more business-focus, less technical
- **Analytics Analyst / Insights Analyst** — common in marketing and product
- **Marketing Analyst / Product Analyst** — domain-specific analyst roles
- **Operations Analyst / Financial Analyst** — function-specific analyst roles
- **Data Specialist / Data Coordinator** — often entry-level data roles
- **Reporting Analyst / BI Analyst** — dashboard and reporting focused

Top Industries Hiring GDA Certificate Graduates

Industry	Why They Hire	Common Tools Used
Technology	Product analytics, growth, user behavior	SQL, Python, Looker, Amplitude
Healthcare	Patient outcomes, cost analysis, operational efficiency	SQL, Tableau, Excel
Finance & Banking	Risk modeling, fraud detection, customer segmentation	SQL, Python, SAS
Retail & E-Commerce	Customer behavior, inventory, campaign ROI	SQL, Tableau, GA4
Consulting	Client projects across industries	Excel, Power BI, SQL
Government / Nonprofit	Policy analysis, program evaluation	Excel, R, Tableau Public
Marketing Agencies	Campaign performance, attribution modeling	GA4, SQL, Tableau

What Employers Want Beyond the Certificate

- **Python:** The most commonly requested skill not in the GDA certificate. Even basic Pandas and Matplotlib proficiency is highly valued.
- **Advanced SQL:** Window functions, CTEs, query optimization, and working with very large datasets.
- **Communication skills:** The ability to explain technical findings clearly to non-technical stakeholders is the #1 soft skill employers mention.
- **Domain knowledge:** Understanding the industry you're applying to (healthcare billing, e-commerce funnels, financial metrics) sets you apart.
- **Portfolio quality:** 2–3 polished projects consistently outperform a list of certifications in hiring decisions.
- **GitHub presence:** Recruiters increasingly search GitHub to verify coding claims on resumes.

A4 — Tools & Software Reference

Quick reference for all four primary tools in the GDA certificate — where to access them, what the free tier gives you, and the specific features most tested in the program.

Google Sheets

- **Access:** sheets.google.com — free with any Google account. No installation.

- **Key features tested:** VLOOKUP, SUMIF, COUNTIF, AVERAGEIF, pivot tables, sorting, filtering, data validation, TRIM, CLEAN.

- **Shortcuts to know:** Ctrl+Shift+L = toggle filter. Ctrl+D = fill down. Ctrl+; = insert today's date. F4 = toggle absolute/relative reference ($).

- **Practice data:** Google provides sample datasets in Course 2 and 4. Also try importing any CSV from Kaggle.

- **Tip:** Use the Explore button (bottom-right of sheet) for instant AI-powered summaries and chart suggestions of your selected data.

BigQuery (SQL)

- **Access:** console.cloud.google.com — BigQuery Sandbox is free with no credit card required. $0 for the first 1TB of query processing per month.

- **Key features tested:** SELECT/FROM/WHERE, GROUP BY, HAVING, all JOIN types, subqueries, CAST, TRIM, COALESCE, CASE statements, date functions.

- **BigQuery-specific syntax:** Use backticks for table names with hyphens: \`project.dataset.table\`. String comparison is case-sensitive.

- **Practice:** BigQuery has free public datasets. Search for "bigquery-public-data" — it includes Chicago taxi trips, New York Citi Bike, and more.

- **Tip:** Always use LIMIT when exploring — SELECT * FROM a billion-row table without LIMIT will process a lot of data. Preview with "Preview" tab instead.

Tableau Public

- **Access:** public.tableau.com — 100% free. Download Tableau Public desktop app or use the web editor.

- **Key features tested:** Connecting to data (CSV, Google Sheets), building worksheets, dimensions vs. measures, drag-and-drop chart building, filters, calculated fields, dashboards.

- **Limitations of Public vs. Desktop:** Tableau Public saves workbooks publicly — do not upload sensitive data. No private saves.

- **Practice:** Tableau provides sample superstore dataset built in. Start there to learn the interface before connecting your own data.

- **Tip:** Use "Show Me" (top-right panel) to see which chart types Tableau recommends based on the fields you've selected.

R & RStudio

- **Access:** Download R from cran.r-project.org and RStudio from posit.co — both free. Or use Posit Cloud (posit.cloud) — browser-based, no install needed.

- **Key features tested:** install.packages(), library(), tidyverse, dplyr verbs (filter/select/mutate/arrange/summarize), pipe operator %>%, ggplot2 layers, R Markdown.

- **Key packages:** tidyverse (installs ggplot2, dplyr, tidyr, readr all at once). That's all you need for Course 7.

- **Reading data:** read_csv("file.csv") from readr. Use read.csv() for base R (also fine).

- **Tip:** Run code line-by-line with Ctrl+Enter while learning. Once confident, run the full script with Ctrl+Shift+Enter. The Environment pane (top-right) shows all objects in memory.

A5 — Recommended Next Steps After Passing

You've passed the Google Data Analytics Certificate. Congratulations — now the real work begins. Here's the roadmap that successful graduates follow to go from "certificate holder" to "employed data analyst."

Month 1–2: Build Your Foundation

- **Complete your capstone write-up** if you haven't already. Use the templates in Chapter 8.
- **Set up GitHub.** Create an account, push your capstone code, and write a README that explains the project clearly.
- **Update LinkedIn** with your certificate, new skills (SQL, Tableau, R, BigQuery), and the capstone project.
- **Start your second portfolio project.** Pick a public dataset from Kaggle or a government portal. Choose a topic you genuinely find interesting — your enthusiasm will show in the write-up.
- **Learn Python basics.** Python is the single most impactful skill to add after the GDA certificate. Complete the free Python for Everybody course (py4e.com) or Kaggle's free Python course.

Month 3–4: Deepen Your Skills

- **Advanced SQL:** Learn window functions (ROW_NUMBER, RANK, LAG/LEAD), CTEs (WITH clause), and query optimization. Mode Analytics SQL tutorial is excellent and free.
- **Third portfolio project.** Aim for something in the industry you want to work in — if healthcare interests you, find a healthcare dataset.
- **Tableau Desktop Specialist.** If you enjoyed Course 6, this certification validates your Tableau skills and is well-recognized by employers.
- **Apply for jobs while learning.** Don't wait until you feel "ready" — the job search takes 3–6 months. Start applying now and treat rejections as practice.

Month 5+: Specialization & Job Search

- **Choose a direction:** Advanced Analytics (Python, statistics, ML → Google Advanced Data Analytics Certificate) or Business Intelligence (dashboarding, data modeling → Google BI Certificate).
- **Informational interviews.** Message 5–10 data analysts on LinkedIn with personalized notes asking for 20-minute calls. Most people are willing to help. These conversations lead to referrals.
- **Join the community.** r/dataanalysis, DataTalks.Club, local analytics meetups on Meetup.com. Visibility in these communities leads to unexpected opportunities.
- **Freelance to build experience.** Upwork, Fiverr, and local nonprofits often need data help. Even one paid project adds "client work" to your resume.

Complementary Certifications Worth Pursuing

Certification	Provider	Time	Why It Helps
Google Advanced Data Analytics	Coursera / Google	~6 months	Adds Python, statistics, ML — the natural next step
Google Business Intelligence	Coursera / Google	~2 months	Deepens dashboarding and data modeling skills
Tableau Desktop Specialist	Salesforce	~4 weeks prep	Validates Tableau skills from Course 6
Microsoft PL-300 (Power BI)	Microsoft	~6 weeks prep	Opens Power BI jobs (huge in enterprise)
AWS Cloud Practitioner	Amazon	~4 weeks prep	Cloud literacy increasingly expected
SQL Certification (DataCamp)	DataCamp	~2–4 weeks	Validates SQL proficiency with a recognized credentia

A6 — Free Resources & Further Reading

A curated list of the resources that GDA certificate students and the data analytics community consistently recommend. All of these are free or have a substantial free tier.

Practice Datasets

Source	URL	Best For
Kaggle Datasets	kaggle.com/datasets	Largest collection; competitions; community notebooks
Google Dataset Search	datasetsearch.research.google.com	Meta-search across all public repositories
UCI ML Repository	archive.ics.uci.edu	Classic academic datasets; structured and clean
Data.gov	data.gov	US government data across every agency and topic
World Bank Open Data	data.worldbank.org	International economic and development data
NYC Open Data	opendata.cityofnewyork.us	Rich urban datasets; great for beginner projects
FiveThirtyEight Data	github.com/fivethirtyeight/data	Journalistic datasets; great for interesting stories

YouTube Channels

Channel	Focus	Best For
Alex The Analyst	Career path, portfolio projects, SQL/Excel	Entry-level job seekers without STEM degree
StatQuest (Josh Starmer)	Statistics and ML explained visually	Understanding the theory behind analytics
Codebasics	Hands-on data analytics projects	Real project walkthroughs end-to-end
Data School (Kevin Markham)	Pandas, SQL, scikit-learn	Clear explanations with practical examples
Luke Barousse	SQL for data analysts	SQL interview prep and real analyst queries
Corey Schafer	Python, Git, SQL fundamentals	Clean, thorough tutorials for beginners
Chandoo	Excel and Power BI mastery	Spreadsheet and dashboard power users

Free Learning Platforms

- **Kaggle Learn** (kaggle.com/learn) — Free micro-courses in SQL, Python, Pandas, data visualization, machine learning. Bite-sized and practical.

- **Mode Analytics SQL Tutorial** (mode.com/sql-tutorial) — The best free SQL tutorial for analysts. Covers window functions, subqueries, and advanced topics.
- **W3Schools SQL** (w3schools.com/sql) — Quick reference and practice for SQL syntax.
- **Google Analytics Academy** (analytics.google.com/analytics/academy) — Free courses on Google Analytics 4.
- **Tableau Training Videos** (tableau.com/learn/training) — Official Tableau tutorials from beginner to advanced.
- **Posit Cloud** (posit.cloud) — Free browser-based RStudio. Run R without any local installation.
- **DataLemur** (datalemur.com) — SQL and statistics interview questions from real tech companies. Essential for interview prep.

Communities to Join

- **r/GoogleCertificates** — The primary Reddit community for GDA certificate students. Ask questions, find study partners, share wins.
- **r/dataanalysis** — Broader data analytics discussions, career advice, and portfolio feedback.
- **r/SQL** — For SQL-specific questions and practice problems.
- **DataTalks.Club** (datatalks.club) — Free community with Slack, office hours, free courses, and a supportive cohort environment.
- **LinkedIn** — Follow data analysts at companies you admire. Post about your projects and learning journey — visibility leads to opportunities.
- **Kaggle** — Participate in discussions and competitions. Your Kaggle profile is a public portfolio.
- **Local Meetup.com groups** — Search "data analytics" or "data science" in your city. In-person networking still has an outsized impact on career opportunities.

A7 — Sample Interview Questions with Model Answers

These are real interview questions asked at entry-level data analyst interviews. Study the model answers — but make them your own by inserting examples from your actual portfolio projects.

Q: Walk me through your data analytics process from start to finish.

Model Answer: I follow the six-phase process I used in my capstone: Ask (define the business question and stakeholder needs), Prepare (identify and assess data sources), Process (clean and validate the data — documenting every change), Analyze (query and model the data to find patterns), Share (visualize findings in a format appropriate for the audience), and Act (deliver specific, actionable recommendations). In my portfolio project analyzing bike-share membership, this process took me from a vague business problem to three concrete recommendations that the stakeholder team could implement immediately.

Q: What is the difference between a LEFT JOIN and an INNER JOIN?

Model Answer: An INNER JOIN returns only rows where a matching record exists in both tables — if a customer has no orders, they don't appear in the result. A LEFT JOIN returns all rows from the left table plus matching rows from the right table. Customers with no orders still appear, with NULL in the order columns. I use LEFT JOINs when I need to find records that exist in one table but not another — for example, identifying customers who have never purchased.

Q: How do you handle missing or null values in a dataset?

Model Answer: My first step is understanding WHY the nulls exist — are they missing because the data was never collected, or because the value genuinely doesn't apply? For analysis-critical columns, I'll either remove rows with nulls (if there are few), substitute a meaningful default using COALESCE in SQL, or flag them as a separate category. I always document what I did and why in the changelog so anyone reviewing the analysis can understand the decisions.

Q: Tell me about a project where you had to clean messy data.

Model Answer: In my customer churn analysis, I imported data from three systems that each stored dates in different formats — one as YYYY-MM-DD, one as MM/DD/YYYY, and one as a Unix timestamp. I wrote SQL CAST operations to normalize all three to DATE type, used TRIM to fix trailing spaces in customer names that were causing join mismatches, and removed 847 duplicate records that appeared during a data migration. I documented every step in a changelog and ran row count checks before and after each cleaning operation. The final dataset reduced from 48,000 rows to 44,200 clean, usable records.

Q: How would you explain a complex finding to a non-technical executive?

Model Answer: I lead with the conclusion, not the process. Instead of "we ran a regression with 14 variables," I say "customers who receive emails within 3 days of signup are 40% more likely to make a purchase in the first month." Then I show one clear chart that supports that claim. I avoid jargon and frame everything in terms of business impact and next steps. If they want the technical detail, I have it ready — but I don't lead with it.

You've Got This.

Every professional data analyst started exactly where you are right now — learning what a JOIN is, figuring out why their VLOOKUP returns #N/A, and wondering if they'll ever feel confident enough to apply for a job. The ones who made it aren't the ones who were naturally talented. They're the ones who kept going.

Build the portfolio. Write the case studies. Post about your projects on LinkedIn. Apply before you feel ready. The certificate gave you the foundation — now go build something real on top of it.

GDACertPrep.com